Exploiting Visual Context Semantics for Sound Source Localization

Xinchi Zhou[1*] Dongzhan Zhou[1*] Di Hu[2] Hang Zhou[3] Wanli Ouyang[1]
[1]The University of Sydney [2]Gaoling School of Artificial Intelligence,
Renmin University of China [3]Baidu Inc.
{xinchi.zhou1,d.zhou,wanli.ouyang}@sydney.edu.au,
dihu@ruc.edu.cn, zhouhang09@baidu.com

Abstract

Self-supervised sound source localization in unconstrained visual scenes is an important task of audio-visual learning. In this paper, we propose a visual reasoning module to explicitly exploit the rich visual context semantics, which alleviates the issue of insufficient utilization of visual information in previous works. The learning objectives are carefully designed to provide stronger supervision signals for the extracted visual semantics while enhancing the audio-visual interactions, which lead to more robust feature representations. Extensive experimental results demonstrate that our approach significantly boosts the localization performances on various datasets, even without initializations pretrained on ImageNet. Moreover, with the visual context exploitation, our framework can accomplish both the audio-visual and purely visual inference, which expands the application scope of the sound source localization task and further raises the competitiveness of our approach.

1. Introduction

We live in a world surrounded by visual and auditory messages and perceiving such multi-modal data can help us better understand our environment [20, 41, 11, 36]. Humans can spontaneously capture the correlations between the sound and appearance of an object. For example, we can associate the chatting and laughing sounds with a group of people passing by. To simulate the human perception ability, audio-visual intelligent systems have been widely explored in recent years and achieved significant improvements.

Sound source localization is an important task in the audio-visual learning field, which aims at locating the sounding objects with the guidance of audio signals. However, obtaining the delicate annotations of object locations, i.e., segmentation masks and bounding boxes, can be rather expensive, especially for large-scale datasets. To overcome this limitation, many works [5, 33, 29, 22, 8] tackle the localization problem in a self-supervised manner, which utilize the audio-visual correspondences as supervision signals. Therefore, the models can generate decent sound source localization results even without precise location annotations. Despite the success, there still exist some issues to be addressed in the current sound source localization frameworks.

Visual context semantics are important for developing a comprehensive scene understanding but have been neglected in many sound source localization frameworks [5, 31, 33, 22]. We identify two advantages for fully exploiting the rich visual context semantics. First, humans tend to locate the sounding objects by visually searching the environment while an effective search normally stems from analysis of the surroundings. For example, when hearing the birds singing, we will subconsciously look for trees and then find the birds. Thus, the visual context is important for humans to accomplish object localization. Second, the visual semantics are inherently discriminative, as shown in previous works [19, 40], and exploiting the intrinsic discriminative correlations within the visual domain is beneficial for the learning process. By interacting with these distinct visual semantics, the audio-visual collaboration is strengthened and the deep model can develop more comprehensive scene perception.

Our key insight is that the internal visual discrimination should be effectively coordinated with the audio information to maximize the supervision utility. A visual reasoning module is introduced to the sound source localization system to exploit the visual context semantics. The module produces the reasoning maps indicating the distributions of prominent context regions as well as the corresponding visual features within these regions. We carefully design the learning objectives to guide the visual semantics so that the extracted context features develop better adaptation to the localization target. Specifically, the visual semantics are encouraged to exhibit higher similarity with the audio features, making them focus more on the sounding objects.

*Equal contribution.

Furthermore, we adopt a consistency loss to minimize their distribution differences between the audio-visual localization map and the reasoning maps. In this way, the cross-modal information flow is further facilitated, which leads to more robust feature representations and eventually improves the localization performances. We also emphasize that our framework does not rely on pre-training on large-scale datasets, e.g., ImageNet [25], but achieves competitive results by training from scratch.

Additionally, since our visual reasoning module can generate several objectness maps, we can also obtain a localization heatmap from the pure visual input by combining these maps. Experimental results indicate that the pure visual heatmaps still yield satisfactory localization results. Therefore, the application scenario can be extended from multi-modal (audio-visual) inference to single-modal (purely visual) inference, and the requirements of input data are also reduced.

Our contributions can be summarized as follows: (1) We introduce a visual reasoning module to exploit the visual semantics, which overcomes the defects of insufficient use of visual information in many previous frameworks. (2) We devise specific loss functions to guide the visual semantics, which improve the localization performance of the entire framework. Experimental results demonstrate that our method surpasses baseline methods by a notable margin on various datasets. (3) Our framework enables both multi-modal and single-modal inference, which expands the application scope of the sound source localization task. We hope this preliminary exploration could provide a new perspective for the self-supervised visual localization field.

2. Related Work

Audio Visual Learning. Audio-visual learning [49, 43] has attracted wide attention with the success of deep learning in recent years. Many sub-fields have achieved great progress, such as audio-visual representation learning [4, 24, 5, 6, 29], audio-visual generation [14, 28, 48, 18], visual sound separation [13, 47, 46, 15, 12, 16, 7], and so on. The correlations between audio and visual messages in videos provide natural supervision for various audio-visual tasks, which enables training with large-scale unlabelled video data. Arandjelovic et al. [4] train audio and visual networks with the correspondence guidance and find that both networks learn feature representations effectively. Owens et al. [29] jointly model the audio and visual components to predict whether the two inputs are temporally aligned. [21, 3] utilize feature clustering approaches to realize the self-supervised learning of audio and video representations. Our work also adopts the correspondence supervision to train the sound source localization framework while incorporating a visual reasoning module to capture the intrinsic discrimination within the visual domain.

Sound Source Localization. Self-supervised sound source localization approaches are fuelled by leveraging the co-occurrence of audio and visual messages in videos [5, 29, 33, 21, 31, 8, 22, 35, 34]. [5, 29] utilize audio-visual correspondence or temporal synchronization as supervision signals to learn the feature representations. In [33], the authors adopt the predicted score maps to filter the visual features and compute the cross-modal similarities. Qian et al. [31] use CAMs to find the approximate locations of objects in a weakly supervised manner. Zhao et al. [47, 46] propose the 'mix-and-separate' paradigm to simultaneously learn the separation and visual grounding. Hu et al. [22] establish a dictionary to store the object features from different categories and realize the class-aware sounding object localization. [38] propose a negative-free method to address the false negative sampling problem. [37] consider the visual scene information by introducing the external objectness confidence maps from the selective search algorithm [39] as pseudo localization annotations. Since the confidence maps are processed beforehand, they cannot be updated during the training process, which may raise the risk of over-fitting the noisy labels. Conversely, our framework is optimized in an end-to-end manner. [44] also devise a proposal-based paradigm to enhance the audio-visual localization system. [8, 27] attempt to mine intra-frame hard samples from the audio-visual localization maps. However, their approaches do not directly investigate the visual domains but focus on the cross-modal associations, which are relatively implicit. In contrast, our method explicitly discovers the visual context semantics and then utilizes the audio signals to provide guidance. In this way, our approach can leverage both the discriminative nature of visual scenes and the cross-modal supervisions, which lead to better localization performances.

Visual Contexts in Multi-modal Learning. Visual context semantics have been explored in many computer vision tasks as the utilization can normally boost the performance of deep models [42, 10, 45, 26]. In addition to the pure vision field, visual context semantics also play an important role in multi-modal learning as they can promote the development of comprehensive scene perceptions. Chatterjee [7] et al. leverage the visual structures as a graph to provide better guidance for visual sound separation. [32] propose the TriBERT framework to accomplish the contextual feature learning across three modalities. Shi et al. [37] introduce visual attention maps from selective search [39] as pseudo localization annotations to the sound source localization task but these external visual messages require additional pre-processing and cannot be updated with the network parameters during training. Our method employs a visual reasoning module to explicitly exploit the visual context semantics and adopts specific constrains to promote the

audio-visual interactions, which effectively improve the localization performances. Moreover, apart from the regular audio-visual localization inference, our reasoning module enables pure visual inference, which expands the application scopes and potentially provides new solutions for the visual localization subject.

3. Methodology

3.1. Overall Framework

The objective of the sound source localization task is spatially localizing the sounding objects with the guidance of the corresponding audio cues. The overall framework is presented in Fig. 1.

For a given audio-visual pair, we can obtain both the audio-visual correspondence map $\mathbf{S_{av}}$ and the reasoning maps $\mathbf{S_{rea}}$ derived from the visual features. During the training phase, we use $\mathbf{S_{av}}$ to compute the audio-visual correspondence (AVC) loss, which requires the framework to distinguish between the positive (correlated) and negative (not correlated) audio-visual pairs. The reasoning maps $\mathbf{S_{rea}}$ are responsible for explicitly discovering the visual context semantics under the supervision of the audio-visual context loss. To further strengthen the audio-visual interactions, we impose a cross-map consistency loss to minimize the distribution differences between the visual reasoning map and the audio-visual correspondence map.

During the testing stage, if the audio messages are available, we can implement the regular sound source localization inference, which utilizes the audio-visual correspondence map to locate the sounding objects. However, the audio cues are not always available. We argue that our approach also works with the situation when we only have single images or the audio tracks are corrupted in videos. Under this circumstance, we adopt the aggregated visual reasoning maps to replace the original audio-visual correspondence map. Experimental results indicate that purely visual testing can achieve on-par performances compared with audio-visual testing (please refer to Sec. 4.5). This advantage relaxes the requirements for the input data and thus expands the application range of the sound source localization task.

3.2. Audio-Visual Correspondence Learning

Given an arbitrary audio-visual pair $\{a^i, v^i\}$ from the video clip i, we aim to find which region in v^i has the highest correlation score with a^i. Thus, we feed v^i into the visual network to extract the visual feature $\mathbf{V^i} \in \mathbb{R}^{C\times H\times W}$, where C, H, W refer to channel, height, width, respectively. Here we omit the batch index for simplicity. The audio input a^i is also fed into the audio network to extract the audio feature $\mathbf{A^i} \in \mathbb{R}^C$. By multiplying the visual feature and the audio feature along the channel dimension, we can obtain the predicted correspondence heatmap $\mathbf{S^{ii}_{av}} \in \mathbb{R}^{H\times W}$, as shown in Eq. 1.

$$S^{ii}_{av}(x, y) = \sum_c V^i(c, x, y) \times A^i(c), \tag{1}$$

where x, y denote the coordinate on the $H \times W$ plane. Please note that both the audio and visual features are normalized along the C dimension before computing the correspondence heatmap.

The predicted heatmap $\mathbf{S^{ii}_{av}}$ denotes the correspondence score for each pixel of the visual feature. However, we need to acquire the overall correspondence score that represents the final decision (corresponding or not). Thus, we apply the global max-pooling on the correspondence heatmap to generate the correspondence score s^{ii}_{av}.

Empirically, the global max-pooling will lead to better performances compared with the global average-pooling operation. One possible reason is that many pixels are invalid or less informative and conducting the average-pooling operation on the whole score map will inevitably introduce these noisy pixels. Thus, the supervision signals are weakened and ultimately cause sub-optimal results. Conversely, the max-pooling operation can suppress the noisy pixels and only retain the values with the highest response, which will benefit the training process.

The correspondence loss requires the network to discriminate the positive and negative pairs at the sample level. If the audio and visual inputs come from the same video clip, then they are labeled as positive pairs, otherwise negative. In practice, the negative pairs are constructed within the entire mini-batch during training. In this way, the model can get access to more diverse negative examples and receive stronger supervision signals. The learning objective is defined as follows:

$$L_{avc} = -\frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \log \left[\frac{\exp\left(s^{ii}_{av}\right)}{\sum_{j=1}^{\mathcal{B}} \exp\left(s^{ij}_{av}\right)} \right], \tag{2}$$

where $\mathcal{B}$ refers to the batch size during training, s^{ij}_{av} refers to the correspondence score between v^i and a^j inside the mini-batch. The correspondence loss encourages the network to increase the similarity between the positive pairs while suppressing that between the negative pairs.

3.3. Exploiting Visual Context Semantics

Audio-visual correspondence learning mainly focuses on the global correlations between audio and visual pairs. Although the global max-pooling can effectively suppress the less informative pixels, some useful pixels may also be removed when computing the overall correspondence score. Chen et al. [8] take this issue into account and firstly incorporate the hard negative regions into the sound source localization problem. We argue that the intrinsic discriminative property within the visual domain has still not been fully

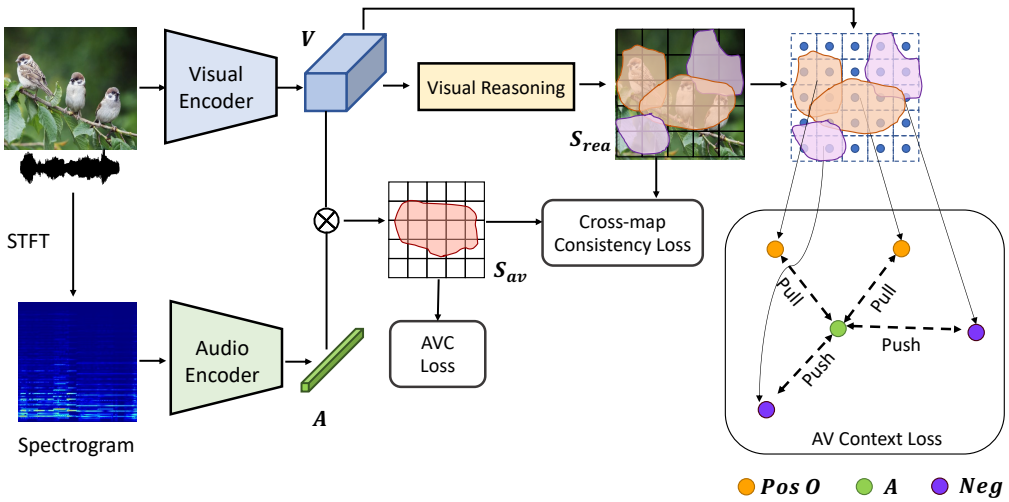

Figure 1. Overview of the whole sound source localization framework. Given an audio-visual pair, we can obtain both the audio-visual correspondence map $\mathbf{S_{av}}$ and the reasoning maps $\mathbf{S_{rea}}$ derived directly from the visual features. In addition to the regular audio-visual correspondence (AVC) loss, we propose the audio-visual context loss and the cross-map consistency loss to guide the exploitation of visual context semantics, which strengthen the audio-visual interactions and bring more robust feature representations. The video index is omitted for simplicity.

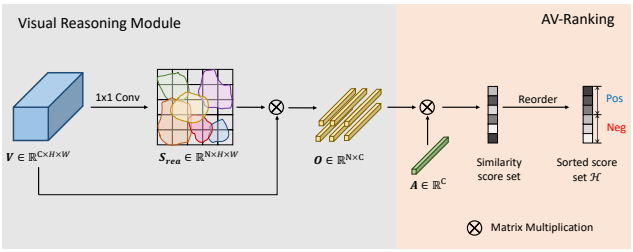

Figure 2. Illustration of the visual reasoning module and the av-ranking strategy. The visual feature $\mathbf{V}$ passes through a 1×1 convolution layer to generate the reasoning maps $\mathbf{S_{rea}}$, which indicate the distribution of possible object locations. The labels (pos/neg) of the reasoning maps are determined by the similarity between the region features $\mathbf{O}$ and the corresponding audio feature $\mathbf{A}$. The video index is omitted for simplicity.

exploited since the sample selection is based on the audio-visual correlation map. Alternatively, we propose a visual reasoning module to simultaneously leverage the natural visual structures and the cross-modal associations.

3.3.1 Visual Reasoning Module Structure

The structure of the reasoning module is illustrated in the left part of Fig. 2. In our reasoning module, the first step is to project the visual feature $\mathbf{V}$ into the visual discriminative space, where the model learns to automatically find the meaningful regions in the images. The feature projection is accomplished by a convolution layer with kernel size $N \times C \times 1 \times 1$ and the outputs are a set of reasoning maps $\mathbf{S_{rea}} \in \mathbb{R}^{N\times H\times W}$, where N refers to the number of selected regions. The convolution kernel can be regarded as a group of learnable projection weights.

As the reasoning maps $\mathbf{S_{rea}}$ indicate the distribution of the possible object locations, we apply the reasoning maps as weights to sum the visual feature $\mathbf{V}$ to obtain the features of the selected regions, denoted as $\mathbf{O} \in \mathbb{R}^{N\times C}$:

$$O(n,c) = \sum_{h,w} V(c,h,w) \times S_{rea}(n,h,w) \tag{3}$$

Although the reasoning mechanism can help utilize the visual contexts, we argue that directly adding this module to the existing framework does not work effectively. Instead, it is necessary to employ specially designed losses to provide clear guidance for the discovered visual semantics and further enhance the cross-modal interactions.

3.3.2 Learning Objectives of Visual Reasoning

The regions selected by the reasoning module may contain both the foreground and background area, which will be discriminated via an av-ranking strategy with the guidance of audio cues, as shown in the right part of Fig. 2. If not specified, we omit the video index i and manipulate visual and audio features from the same video. By conducting the dot-product between the region features $\mathbf{O}$ and the corresponding audio feature $\mathbf{A}$, we can get a set of the similarity score $\{h_k|k=1,2,...N\}$.

The similarity scores are then sorted in descending order to generate the reordered score set $\mathcal{H}$ and the corresponding index set $\mathcal{I}$:

$$\begin{cases} \mathcal{H} = \text{sorted}(\{h_1, h_2, ..., h_N\}), \\ \mathcal{I} = \text{argsort}(\{h_1, h_2, ..., h_N\}) \end{cases} \tag{4}$$

The first N_P values in $\mathcal{H}$ are considered positive subset while the last N_Q values negative subset. The strategy to distinguish the positive and negative regions provides higher flexibility during training since it encourages the model to independently discover the cross-modal interaction patterns, which may lead to stronger feature representations. The positive score P and negative score Q can be computed by averaging the values in the positive subset and negative subset, respectively. We can then define the audio-visual context loss $L_{context}$ in Eq. 5, which is of a similar form as Eq. 2.

$$L_{context} = -\frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \left[\log \frac{\exp\left(P_i\right)}{\exp\left(P_i\right) + \exp\left(Q_i\right)} \right], \quad (5)$$

where P_i and Q_i refer to the average positive and negative score, respectively, for the i^{th} matched audio-visual pair inside the mini-batch.

Both the reasoning maps $\mathbf{S_{rea}}$ and the AV-correspondence map $\mathbf{S_{av}}$ indicate the potential locations of target objects but they come from different feature sources, i.e., $\mathbf{S_{rea}}$ via visual context semantics and $\mathbf{S_{av}}$ via audio-visual correlations. To promote the information interactions between the two sources, we devise a consistency constraint on these two types of localization heatmaps. Before computing the loss value, the negative fragments need to be filtered from the reasoning maps. We take the first N_P indexes from the index set $\mathcal{I}$ and utilize these *positive* indexes to pick up the maps representing the sounding objects. By taking the average value at each pixel on the N_P foreground maps, we acquire the aggregated map $\mathbf{S}^{'}_{\mathbf{rea}} \in \mathbb{R}^{H \times W}$, which is of the same shape as $\mathbf{S_{av}}$.

We implement the consistency constraint by minimizing the distribution differences between the aggregated reasoning map and the AV-correspondence map, as formulated below:

$$L_{div} = \mathcal{D}_{JS}(\mathbf{S}^{'}_{\mathbf{rea}} \,||\, \mathbf{S_{av}}), \quad (6)$$

where $\mathcal{D}_{JS}$ refers to the Jensen-Shannon Divergence.

3.3.3 Final Objective

The final objective of the entire sound source localization framework consists of all the losses for audio-visual correspondence and visual context semantics exploitation.

$$L = L_{avc} + \lambda_1 L_{context} + \lambda_2 L_{div}, \quad (7)$$

where λ_1 and λ_2 are loss weights for balancing the importance of the learning objectives.

4. Experiments

4.1. Implementation Details

The framework is trained using the Adam optimizer [23] with a learning rate of 1e-4 and batch size of 128. The weight decay is set to 1e-4. The loss weights λ_1 and λ_2 are set to 0.1. All experiments are implemented with the PyTorch framework [30] on 4 NVIDIA 2080TI GPUs. Details about model configuration and data processing are provided in the supplementary.

4.2. Datasets and Evaluation Metrics

VGGSound. The VGGSound dataset [9] is a recently released audio-visual dataset, which consists of 10-second video clips crawled from YouTube. We train our models on the subset of 220 categories and evaluate the results on the VGG-SS test set [8].

VGGSound-MI. We create another subset from the VGGSound dataset for training and evaluation, which contains around 30k clips from 39 musical instruments. This subset simulates the situation where the data amount is smaller and the category distribution is more concentrated. The category list will be provided in the supplementary.

Flickr SoundNet. This dataset was firstly proposed in [6], containing over 2 million video clips from Flickr. Following previous works [33, 31, 8], we adopt the human-annotated subset for quantitatively evaluation. In our setting, we randomly sample 10k video clips for training and evaluate on 250 annotated pairs.

AudioSet. AudioSet is another large-scale audio-visual dataset proposed in [17]. We adopt the subset of around 50k video clips spanning 15 musical instruments. The videos from the 'unbalanced' split are used for training and those from the 'balanced' split are used for testing. More details are provided in the supplementary.

Evaluation Metrics. Following previous works [33, 8, 35], we adopt the Consensus Intersection over Union (CIoU) and Area Under Curve (AUC) as the evaluation metrics, which are calculated with the predicted sounding object locations and the Ground-Truth bounding boxes.

4.3. Quantitative Results

In this section, we compare our method with recent sound source localization methods on different datasets. The results on the VGGSound and VGGSound-MI are summarized in Table. 1. From the results, we can see that our method consistently surpasses all the competing frameworks by a notable margin on different metrics, which demonstrate the effectiveness of explicitly exploiting the richer visual context semantics. Specifically, for the VGGSound dataset, the AUC increases from 0.366 to 0.376 (+1.0%) and the CIoU@0.5 increases from 0.322 to 0.350 (+2.8%); for the VGGSound-MI subset, the AUC increases

Method	VGGSound		VGGSound-MI	
	CIoU@0.5	AUC	CIoU@0.5	AUC
Attention [33]	0.185	0.302	0.243	0.335
DMC [21]	0.193	0.286	0.270	0.362
AVobject [2]	0.297	0.357	0.339	0.382
LCBM [35]	0.322	0.366	0.347	0.392
LVS [8]	0.303	0.364	0.333	0.389
Ours	**0.350**	**0.376**	**0.365**	**0.402**

Table 1. Quantitative results on the VGGSound and the VGGSound-MI datasets. CIoU@0.5 means that the IoU threshold for the CIoU metric is 0.5. LCBM [35] is a weakly-supervised framework as it uses category labels during training while other methods are trained in a self-supervised manner.

from 0.392 to 0.402 (+1.0%) and the CIoU@0.5 increases from 0.347 to 0.365 (+1.8%). The results also verify that the improvements are robust with different data amount and category distributions. Moreover, our framework is trained from scratch without using any category labels, further indicating the benefits of our method.

Among all the competing methods, the work most similar to us is LVS [8] as the background regions are explicitly considered in the supervision. However, the results indicate that our method with the visual reasoning module exhibits better localization performances. We suspect that the possible reasons are as follows. First, LVS only relies on the audio-visual contrastive modes while our method simultaneously utilizes the intrinsic discriminative attributes within the visual domain and the cross-modal associations. These two sources can serve as complements for each other and provide more diverse supervision signals. Second, the cross-modal synergy is enhanced with the employment of the visual reasoning module and the relevant losses. By capturing the audio-visual interactions more reasonably, the model can learn stronger feature representations from the multi-modal semantics and raise the overall localization performance.

We also conduct experiments on the Flickr-SoundNet and AudioSet subset to examine the adaptability of the model on different datasets. The results on the Flickr SoundNet are summarized in Table. 2. We can see that our model outperforms the competitive LVS [8] method when training on the VGGSound or Flickr SoundNet dataset. For the AudioSet subset, compared with the baseline method, our approach achieves +2.1% and +3.0% gains on AUC and CIoU@0.5, respectively. The results further demonstrate the versatility of our model on different datasets.

4.4. Ablation Study

In this part, we perform ablation experiments to investigate the effects of various factors proposed in our approach.

Model	Training Set	Testing Set	CIoU@0.5	AUC
LVS [8]	VGGSound	Flickr	0.651	0.551
Ours			0.775	0.596
LVS [8]	Flickr10k	Flickr	0.582	0.525
Ours			0.631	0.551

Table 2. Experimental results on Flickr-SoundNet dataset. We conduct training on both VGGSound and Flickr-SoundNet (with 10k samples) dataset.

$L_{context}$	Ra	L_{div}	CIoU@0.5	AUC
✗	✗	✗	0.307	0.354
✓	✗	✗	0.329	0.362
✓	✓	✗	0.340	0.370
✓	✓	✓	0.350	0.376

Table 3. Ablation study. All the models are trained on the VGGSound dataset and evaluated on the VGG-SS test set. We explore the influences of different learning objectives and the av-ranking strategy (denoted as Ra). The results indicate that each proposed module contributes to the performance gains.

4.4.1 Effect of the av-ranking strategy

For the regions predicted by the reasoning module, we propose an av-ranking strategy to divide them into positive and negative areas based on the similarities with the corresponding audio vectors. To strip this factor, we replace the av-ranking strategy by naively tagging the region features according to the order, e.g., directly specifying the first several maps as positive and the remaining ones negative. From the results in Table. 3, we can observe that the av-ranking strategy exhibits better performance, which may come from flexibility during training. Specifically, the model can freely find the most reasonable feature matching solution for each audio-visual pair, instead of forcing the visual semantics to follow a certain order. Thus, the visual network can perceive the audio cues more effectively and hence boost the localization accuracy.

4.4.2 Impacts of different learning objectives

We investigate the impacts of the learning objectives in our approach, as illustrated in Table. 3. Training only with the AV-correspondence loss serves as the naive baseline method. We notice that using $L_{context}$ still outperforms the baseline method even without the av-ranking strategy, which demonstrates the advantages of explicitly incorporating the visual context semantics. Additionally, the av-ranking strategy provides a more rational training mechanism and hence amplifies this positive effect. By minimizing the distribution gaps between the audio-visual correspondence map and the visual reasoning map via the divergence loss L_{div}, we build a new bridge to connect the two modalities. With more efficient audio-visual interactions, the model can acquire more comprehensive information and

Method	CIoU@0.5	AUC
Baseline	0.307	0.354
Vanilla	0.301	0.350
Ours	0.350	0.376

Table 4. Comparison of the vanilla reasoning and our method on the VGG-SS test set. Baseline represents the original network while 'Vanilla' refers to simply adding the visual reasoning module to the original network to augment the visual features. Both methods are trained only with L_{avc}.

thus learn more robust feature representations, which may explain the further gains when applying the divergence loss. Overall, we can observe that all the proposed learning modules contribute to the performance improvements.

4.4.3 Compared with vanilla reasoning

Although the visual reasoning module can utilize the visual context semantics, merely adding a reasoning structure to the existing backbone is not enough and the specially designed supervisions are necessary. Table. 4 shows the comparison between adopting the reasoning module with *no* explicit supervisions (denoted as 'Vanilla') and our approach. We can see that vanilla reasoning performs even slightly worse than the baseline approach, probably due to the incompatibility of the unconstrained features to the localization task. Therefore, the improvements cannot be simply realized by employing the reasoning module, but our specially designed learning objectives can effectively raise the localization performance.

4.5. Image-only Inference

By explicitly utilizing the visual context semantics during training, our framework enjoys an inherent advantage of image-only inference. In other words, our model can also work in the situation when the audio tracks are not available in videos or the current data are in the format of individual images, which expands the application scope of the sound source localization task.

When conducting the image-only inference, we adopt the aggregated reasoning maps to replace the original AV-correspondence map to predict the object locations. Since audio cues are not available currently, we apply the pooling operation on all the reasoning maps to accomplish the combination, instead of merging the foreground maps as in Eq. 6. The rational for combining all the reasoning maps is that the responses of the negative regions will gradually decrease during training so that the positive semantics can dominate the aggregated reasoning map.

The results of the image-only inference are shown in Table. 5, where we can see the single-modality inference can achieve on-par or even slightly better performances compared with the audio-visual counterpart. To our knowledge, this appealing property has not appeared in previous sound source localization works. The results further demonstrate the reliability of the learned visual context semantics.

	VGGSound		VGGSound-MI	
Method	CIoU@0.5	AUC	CIoU@0.5	AUC
Baseline	0.307	0.354	0.326	0.381
Ours (audio-visual)	0.350	0.376	0.365	0.402
Ours (image-only)	0.352	0.378	0.372	0.404
AV-detector [1]	-	-	0.369	0.398

Table 5. Results under the image-only inference scenario on the VGGSound and VGGSound-MI datasets. The image-only inference can achieve on-par or even better performances compared with the audio-visual counterparts.

In [1], the authors extract pseudo bounding box annotations from the AV-localization heatmaps and adopt the annotations to train an object detector. Thus, the framework also accomplishes the transition from audio-visual localization to uni-modal localization as the detector can directly infer the images. Since the codes and models are not released currently, we re-implement their method and report the results in Table. 5. We argue that the detector training relies on carefully selecting hyper-parameters and consumes additional computation resources. In this way, it may not be appropriate to directly compare these two approaches since our method can realize the image-only inference without the subsequent training or additional framework. This inherent advantage provides a new feasible solution to the localization only in the visual domain. The advantage of [1] is that the pre-trained detector can identify instances in images and object categories, which cannot be achieved by our model. We hope that this attribute can be combined with our framework in the future.

4.6. Qualitative Results

We visualize the AV-localization maps on the VGG-SS test set and compare them with the LVS [8] method, as shown in Fig. 3. The results indicate that our approach enjoys better localization ability. We can also see that our method is still able to generate fair localization predictions even under relatively difficult scenarios, such as the python in camouflage (bottom right of Fig. 3), while LVS produces wider results of the entire frame. The visualization results further validate the effectiveness of the sufficient utilization of visual context semantics. More visualization examples are provided in the supplementary materials.

5. Discussion and Future Works

In addition to the performance improvements, we observe that explicitly exploiting the visual contexts also contributes to the training process. Fig. 4 illustrates the AUC values at different epochs during training, where we can see that our approach consistently outperforms the baseline through the entire process. We speculate that utilizing

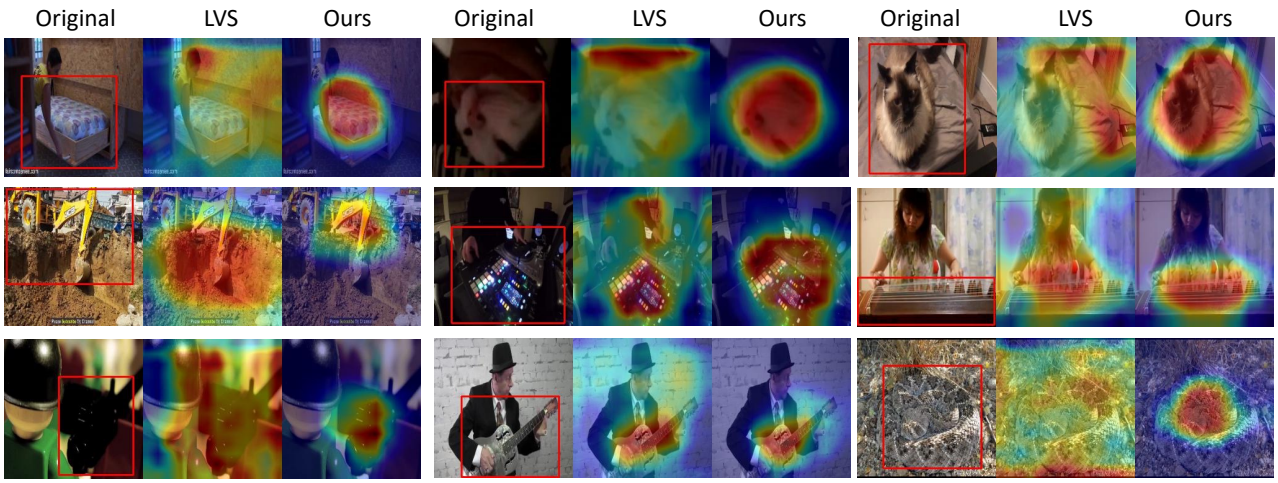

Figure 3. Qualitative results of sound source localization on VGG-SS test set compared with LVS [8].

the visual contexts can bring more diverse supervisions and hence promote the model divergence. The stability of training also reflects the robustness of our method.

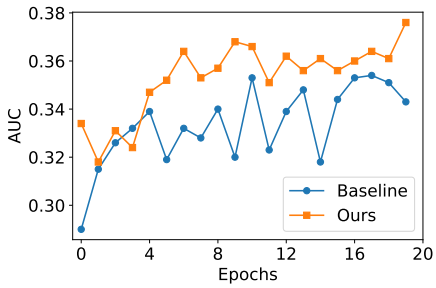

Figure 4. AUCs at different epochs during training. We compare the performances between our approach (with visual semantics exploitation) and the baseline approach.

The experiments have demonstrated the advantages of harvesting the visual context semantics in the sound source localization task. However, the simple visual reasoning module employed in our framework is just one feasible solution. In the future, we will investigate more diverse architectures to accomplish the visual semantics extraction, such as multi-level feature fusion, feature pyramid, etc. Furthermore, the importance of proper learning objectives to enhance the cross-modal interaction should also not be neglected.

Our method proves that leveraging the abundant videos of unconstrained scenes to realize the self-supervised purely visual localization is a feasible way, as our model achieves on-par or even better performance under the image-only inference compared with the audio-visual counterpart. Despite this exciting discovery, there still exist many issues to be addressed, such as object discrimination. Specifically, the predicted heatmaps only indicate the object locations and do not contain category information. [22] learn to discriminatively localize sounding objects but the inference still relies on the audio information. The reasonable way to combine the category discrimination and purely visual inference capabilities still needs exploration. Moreover, discrimination at the instance level for multi-object scenes is also a worth considering problem.

6. Conclusion

In this work, we delve into the exploitation of the visual context semantics in the sound source localization task, which overcomes the problem of the insufficient use of the visual context cues in many previous works. We carefully design the learning objectives that can provide stronger guidance for the extracted visual semantics while strengthening the audio-visual interactions. Experimental results indicate that our approach can effectively boost the sound source localization performances on various datasets. Moreover, since the model is instructed to explicitly mine the visual semantics during training, our framework can realize both the multi-modal (audio-visual) and the single-modal (image-only) inference, where the two inference types achieve similar performances. This unique advantage expands the application field of sound source localization and potentially brings a new direction for the self-supervised visual localization subject.

Acknowledgement

Wanli Ouyang was supported by the Australian Research Council Grant DP200103223, Australian Medical Research Future Fund MRFAI000085, CRC-P Smart Material Recovery Facility (SMRF) – Curby Soft Plastics, and CRC-P ARIA - Bionic Visual-Spatial Prosthesis for the Blind. Di Hu was supported by the National Natural Science Foundation of China (NO.62106272) and the Young Elite Scientists Sponsorship Program by CAST (2021QNRC001).

References

[1] Triantafyllos Afouras, Yuki M Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. Self-supervised object detection from audio-visual correspondence. *arXiv preprint arXiv:2104.06401*, 2021.

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

[3] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020.

[4] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[5] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018.

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.

[7] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1204–1213, 2021.

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

[9] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[10] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019.

[11] Francesca Frassinetti, Nadia Bolognini, and Elisabetta Làdavas. Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental brain research*, 147(3):332–343, 2002.

[12] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020.

[13] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018.

[14] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019.

[15] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888, 2019.

[16] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE, 2021.

[17] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[18] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019.

[19] Robert Harb and Patrick Knöbelreiter. Infoseg: Unsupervised semantic image segmentation with mutual information maximization. In *DAGM German Conference on Pattern Recognition*, pages 18–32. Springer, 2021.

[20] Nicholas P Holmes and Charles Spence. Multisensory integration: space, time and superadditivity. *Current Biology*, 15(18):R762–R764, 2005.

[21] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019.

[22] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33:10077–10087, 2020.

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

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

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[26] Xudong Lin, Lin Ma, Wei Liu, and Shih-Fu Chang. Context-gated convolution. In *European Conference on Computer Vision*, pages 701–718. Springer, 2020.

[27] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Unsupervised sound localization via iterative contrastive learning. *arXiv preprint arXiv:2104.00315*, 2021.

[28] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Ma-

tusik. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7539–7548, 2019.

[29] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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

[32] Tanzila Rahman, Mengyu Yang, and Leonid Sigal. Tribert: Full-body human-centric audio-visual representation learning for visual sound separation. *arXiv preprint arXiv:2110.13412*, 2021.

[33] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[34] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4863–4867. IEEE, 2022.

[35] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3308–3317, 2022.

[36] Ladan Shams and Robyn Kim. Crossmodal influences on visual perception. *Physics of life reviews*, 7(3):269–284, 2010.

[37] Jiayin Shi and Chao Ma. Unsupervised sounding object localization with bottom-up and top-down attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1737–1746, 2022.

[38] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3222–3231, 2022.

[39] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[40] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10052–10062, 2021.

[41] Virginie Van Wassenhove, Ken W Grant, and David Poeppel. Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, 102(4):1181–1186, 2005.

[42] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[43] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022.

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

[45] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12985–12994, 2020.

[46] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019.

[47] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.

[48] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3550–3558, 2018.

[49] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18(3):351–376, 2021.