Unsupervised Sounding Object Localization with Bottom-Up and Top-Down Attention

Jiayin Shi Chao Ma[*]
MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{shijiayin1995, chaoma}@sjtu.edu.cn

Abstract

Learning to localize sounding objects in visual scenes without manual annotations has drawn increasing attention recently. In this paper, we propose an unsupervised sounding object localization algorithm by using bottom-up and top-down attention in visual scenes. The bottom-up attention module generates an objectness confidence map, while the top-down attention draws the similarity between sound and visual regions. Moreover, we propose a bottom-up attention loss function, which models the correlation relationship between bottom-up and top-down attention. Extensive experimental results demonstrate that our proposed unsupervised method significantly advances the state-of-the-art unsupervised methods. The source code is available at https://github.com/VISION-SJTU/USOL.

1. Introduction

As human beings, we can easily locate the sounding objects in visual scenes even without the help of the inherent localization ability of our auditory system. This is because we perceive temporally synchronized visual scenes and their corresponding sounds throughout our entire life and learn the correspondence unconsciously. In contrast, in the context of machine learning, given a pair of image and sound examples, the sound localization task that aims at localizing the sounding objects in the visual scene remains challenging.

In recent years, works about sound localization are mainly based on audiovisual synchronization. They jointly train visual and sound networks to extract deep visual and audio features respectively. Then an integration module fuses the features from the two modalities and is trained on the fused representation to learn the temporal correspondence, thus performing sound source localization [3, 24, 18, 29, 33, 20, 36, 26].

While most recent sound localization methods are limited to musical instruments [3, 36, 19], we focus on the problem of unconstrained visual scenes in this work. In [29], Senocak *et al.* proposed an audiovisual attention mechanism to capture salient regions in unconstrained real-life visual scenes in an unsupervised setting. However, the localization accuracy based on this unsupervised method is not satisfying. To improve the performance, the authors annotated 5k visual-audio samples with bounding boxes to train the model in a supervised way. Several other methods also provide additional supervision. Qian *et al.* [26] leveraged the category labels of images and sounds and established sound-object label alignment. They adopted Class Activation Map (CAM) to measure class-specific correspondence on each spatial grid.

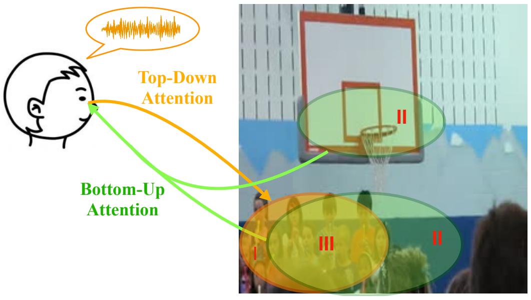

Figure 1. **A brief illustration of bottom-up and top-down attention.** Bottom-up attention is focused more on the basket and the group of people as they stand out from the background. When the subject hears a voice singing, guided by bottom-up attention, his top-down attention can be quickly focused on the group of people.

In summary, unsupervised methods for sound localization in unconstrained real-life visual scenes remain challenging. This limitation derives from the fact that current unsupervised methods learn the audiovisual attention purely from the temporal correspondence. However, when we look for the sound source in a visual scene, the sound information is not the only clue. The visual scene itself also provides meaningful information about where the potential sounding objects can be.

In this paper, inspired by findings about the attention

[*] Corresponding author.

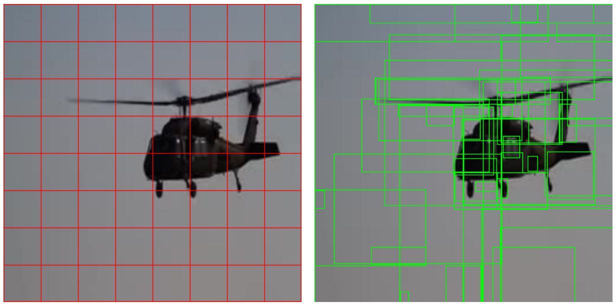
(a) Uniformly scattered state (b) Bottom-up attention

Figure 2. Conventional sound localization methods attend to the equally-sized regions from a scattered state as shown in (a). Our bottom-up attention module allows audiovisual attention map to be generated based on an inherent visual objectness attention map which is shown in (b).

mechanism in cognitive science [21], we introduce an unsupervised method based on the bottom-up and top-down attention mechanism to perform the sound localization task. Figure 1 gives a simple example of the two forms of attention. The subject pays attention to the basket and the group of people at first. Then when the subject hears a singing sound, he is likely to focus more on the group of people, *i.e*., the sound source. More details about top-down attention and bottom-up attention will be discussed in Section 2.

Typically in a sound localization network, visual CNN outputs visual feature maps and sound CNN extracts sound features. The conventional audiovisual attention mechanism encourages the visual features at sound source pixels to take higher similarity with the sound features. However, because CNN features correspond to a uniform grid of equally-sized image regions [2], the module attends to each spatial pixel equally to learn the audiovisual correspondence. This attention mechanism gives little consideration to how likely the image regions would be attended to without the sound information. As shown in Figure 2, the conventional attention mechanism in the sound localization task starts from a uniformly scattered state, which is out of line with the way we perceive the world. Even without any sound, we perceive the silent visual scene with our attention focused on some particular parts as they stand out from the background because of their color, size, or other properties. Hearing the sound, to find the correlation between the sound and the visual scene, our original visual attention is modified to be focused more on sound source areas.

We define the inherent visual attention as bottom-up attention and the audiovisual attention as top-down attention. Our proposed model generates these two attention maps. The bottom-up attention map represents the category-independent objectness score at each spatial grid based on their inherent properties relative to the background. The top-down attention map draws the similarities of deep visual and audio features. Top-down attention map is generated under the guidance of bottom-up attention map.

Specifically, we implement the bottom-up attention module with selective search proposed in [34]. Selective search generates a list of category-independent object regions based on a variety of grouping criteria. Pre-trained object detection models are also tested.

Moreover, to better correlate the two attention maps, we present a bottom-up attention loss function which is modified from the conventional cross entropy loss function. With the cross entropy loss, the two attention maps are trained to be as similar as possible. However, top-down audiovisual attention map is supposed to be guided by bottom-up attention map, but not copy it. Our designed bottom-up attention loss function focuses more on ignoring the inconspicuous area instead of seeing all salient areas. As illustrated in Figure 1, we aim at reducing the area I and maximizing the area III, while we ignore the area II. Experimental results show that it helps improve the localization and sound discrimination ability.

Our method does not require human annotations or category supervision. And our model only needs a 10k size of training set to achieve a new state-of-the-art unsupervised performance. Our supervised implementation using a pre-trained Faster RCNN [28] also achieves the state-of-the-art supervised performance.

In summary, the contributions of our work are three-fold: (1) We propose an unsupervised method for sounding object localization based on the bottom-up and top-down attention mechanism which correlates the visual objectness and audiovisual correspondence; (2) We present a new bottom-up attention loss to describe the guiding relationship of bottom-up and top-down attention; (3) We achieve state-of-the-art results on the public unconstrained sound localization dataset.

2. Related Work

Sound Localization in Visual Scenes. Several approaches have been proposed for sound source localization. Recent methods in visual context mainly focus on joint modeling of audio and visual modalities [3, 24, 18, 29, 33, 20, 36, 26, 19]. [3] performed unsupervised sound localization through learning the audiovisual correspondence in the context of musical instruments. The work of [24] trained a neural network to predict the audiovisual alignment. Tian *et al*. [33] leveraged audio-guided visual attention and temporal alignment to capture semantic regions of sound sources. In [29], the authors proposed an attention mechanism to capture primary areas in an unsupervised way. They also manually annotated a sound source localization dataset of 5k samples from the Flickr-SoundNet dataset [5] for quantitative evaluation of sound localization task and supervised training. Zhao *et al*. [36] and Tian *et al*. [32] employed mixthen-separate frameworks to associate the audio and visual feature maps in the context of musical instruments. The work of [26] adopted CAM to measure class-specific correspondence on each spatial grid. In [19], the authors divided the instrument related datasets [13] [36] into single-source subset and multi-sources subset and then aggregate object localization in single-source videos to build discriminative object representation. [8] proposed automatic negative mining.

We propose an unsupervised method that needs no human annotations and no category labels to perform the sound localization task in unconstrained visual scenes.

Bottom-Up and Top-Down Attention. Our work is motivated by the findings of bottom-up attention and top-down attention in cognitive science and vision science. As our brain has a limitation in its capacity to process massive sensory impressions coming together, attention helps select relevant impressions and ignore irrelevant ones. Currently, there are two commonly distinguished types of attention: bottom-up attention and top-down attention. Bottom-up attention, also called stimuli-driven attention, is purely based on stimuli that are salient because of their inherent properties relative to the background. On the other hand, top-down attention refers to the internal guidance of attention based on prior knowledge, willful plans, and current goals. The two forms of attention are incorporated into a global saliency map [21].

Region Proposal and Object Detection. Region proposal algorithms aim at generating possible object locations for segmentation and detection tasks.

Generating category-independent region proposals methods include objectness [1], selective search [34], category-independent object proposals [10], constrained parametric min-cuts (CPMC) [7], multi-scale combinatorial grouping [4], and Ciresan *et al.* [9]. Selective search [34] combines the strength of both exhaustive search and segmentation. It uses a diverse set of complementary and hierarchical grouping strategies to yield object-class independent region proposals.

Object detection is the task of detecting instances of objects of a certain class within an image. Recently, deep learning techniques [16, 22] have emerged as powerful methods for learning feature representations automatically from data, and provided major improvements in object detection [15, 31, 14, 28, 27]. Faster RCNN [28] is based on work of RCNN [15] and Fast RCNN [14]. It uses a Region Proposal Network to generate a set of proposals and remains one of the best object detection frameworks.

In this paper we use selective search [34] to implement bottom-up attention as it is able to generate good region proposals based on the inherent properties of images without any supervision. We also conduct experiments on object detection methods with Faster RCNN [28] for comparison.

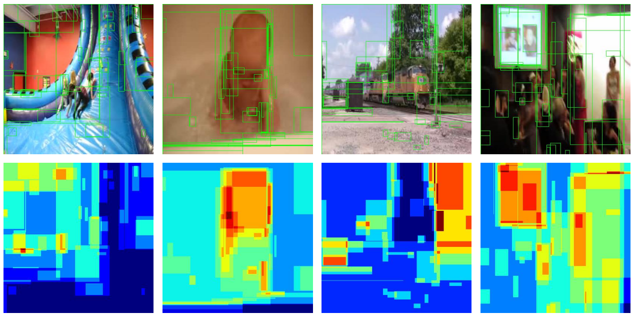

Figure 3. **An illustration of the bottom-up attention map generation process.** Selective search generates region proposals and arranges them in the decreasing order of objectness. Top 50 region proposals shown in the first row are weighted summed to produce the bottom-up attention maps in the second row. The attention maps are normalized for visualization.

3. Proposed Method

In this section, we present our model architecture and proposed attention loss function. The framework is illustrated in Figure 4. In Section 3.1, we describe our approach to implement a bottom-up attention module. In Section 3.2, we present the architecture of the top-down attention model and in Section 3.3, we describe the bottom-up attention loss function to train our network.

3.1. Bottom-up Attention: Objectness at First Glance

Given an RGB image V of size $H \times W \times 3$, the bottom-up attention module generates a $H \times W$ confidence score map A_{bottom} to present the objectness of each pixel before being fed into the deep neural network. This attention map does not involve sound information or object category knowledge and is used as a guidance of top-down attention.

Unsupervised Setting. Specifically, we implement this module using selective search [34]. Selective search [34] firstly over segments the image according to the method described in [12]. Then the algorithm recursively combines the smaller similar regions into larger ones using a diverse set of grouping strategies and thus yields a list of object-class independent region proposals. These proposals are arranged in decreasing order of objectness. We choose the first K regions and then the confidence score map A_{bottom} is calculated as:

$$A_{bottom_{i,j}} = \sum_{k=1}^{K} weight_k * I_{r_{(i,j),k}}, \tag{1}$$

where indicator function $I_{r_{(i,j),k}}$ is defined as:

$$I_{r_{(i,j),k}} = \begin{cases} 1 & \text{if } (i,j) \text{ in region } k, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

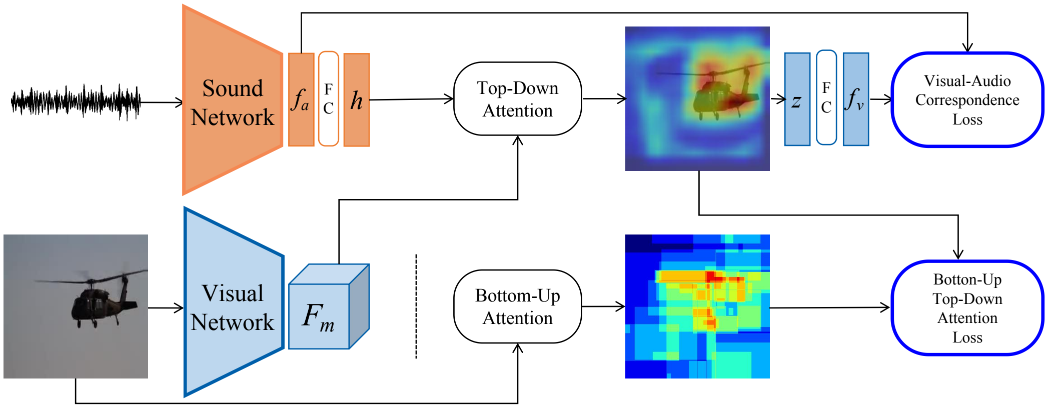

Figure 4. **An overview of our proposed unsupervised learning model based on bottom-up attention and top-down attention.** Bottom-up attention map is generated by the bottom-up attention module from the original image. Visual and audio features are extracted by visual network and sound networks respectively. Then top-down attention is captured by visual-audio correspondence under the guidance of bottom-up attention.

and the $weight_k$ is the objectness weight of region k. The process is illustrated in Figure 3, with K set to 50.

The selective search algorithm considers four types of similarity when combining the initial small starting segmentation regions into larger ones. These similarities are color similarity, texture similarity, size similarity, and fill similarity. Color similarity is measured using the normalized color histogram intersection. Texture similarity is measured using texture histogram derived from fast SIFT. Size similarity encourages small regions to merge early. And fill similarity measures how well two regions fit with each other.

On the one hand, these inherent properties accord with the bottom-up attention in our cognitive system which is purely driven by stimuli. On the other hand, it needs no human annotations for training. Besides, it does not need a pre-train network either. Therefore, we refer to this bottom-up attention implementation with selective search [34] as an unsupervised setting.

Supervised Setting. In contrast to category-independent region proposals, object detection networks locate object instances and determine their classes too. Deep learning techniques have made remarkable breakthroughs in the field of object detection. With many pre-trained models for object detection available, we also implement our bottom-up attention using pre-trained models for object detection.

Specifically we choose Faster RCNN proposed in [28] pre-trained on PASCAL VOC dataset [11]. Given an image V, it generates a list of region proposals. Similarly, we arrange all detected bounding boxes in decreasing order of confidence score and keep those whose confidence scores are larger than a tunable hyperparameter threshold τ. The confidence score map is calculated similarly as defined in Equation 1. The only differences are that here the $weight_k$ is the confidence score generated by Faster RCNN and that region number K is determined by the confidence threshold τ.

As this bottom-up attention implementation with Faster RCNN requires a pre-trained model, we refer to this implementation as a supervised setting. It should be noted that our supervised method still needs no human annotations, which is different from other existing supervised methods in the sound localization task.

3.2. Top-down Attention: Attention Guided by Another Attention

Our visual-audio top-down attention module is similar to work in [29]. Given an audiovisual pair: image V of size $H \times W \times 3$ and raw audio S of size L from an unconstrained video sample, the top-down attention network outputs an attention map A_{top} of size $\lfloor \frac{H}{16} \rfloor \times \lfloor \frac{W}{16} \rfloor$, a visual feature f_v of size 1000 and an audio feature f_a of size 1000.

A VGG16 network proposed in [30] is implemented to extract deep visual features of the input image. With V input into the network, the output feature map of layer conv5_3 of size $\lfloor \frac{H}{16} \rfloor \times \lfloor \frac{W}{16} \rfloor \times 512$ is output as F_m used for further attention calculation.

The SoundNet audio network proposed in [5] is implemented to extract audio features from a 1-D audio signal. We only keep the 1000-D object distribution in conv8. Raw waveform S is input into 1-D CNN and the features are temporally average pooled to get a 1000-D feature f_a. To adapt to the visual features, like [29], f_a is transformed by two fully connected layers to a 512-D feature vector h.

Then A_{top} is calculated as:

$$A_{top_{i,j}} = \bar{F}_{m_{i,j}} \cdot \bar{h}, \tag{3}$$

where $\bar{x}$ denotes the $l2$-normalized vector of x. Then as suggested in [29], attention map A_{top} is softmax normalized.

To give a connection to A_{top} with sound source location, similar to [35, 6, 5], with this top-down attention, visual feature z is obtained by:

$$z = \sum_{i,j} A_{top_{i,j}} \cdot F_{m_{i,j}} \tag{4}$$

Next, the visual feature z is transformed through two fully connected layers to get 1000-D f_v.

At last, the network outputs A_{top}, f_a, f_v. A_{top} is referred to as top-down attention map. It represents the similarity between sound embeddings and visual regions.

3.3. Loss Function

Visual Audio Correspondence Top-down attention is also known as goal-driven attention. We define the goal here in the sound localization task is to learn the correlation of audio and visual features. Similar to [29], we impose that corresponding visual and audio features are close to each other while non-corresponding pairs are far from each other. We use the triplet loss [17] for learning. With visual feature f_v regarded as the query, the corresponding audio feature f_a is the positive sample f_a^+. At each iteration, we randomly select the sound from another sample f_a^- in the training set as a negative sample for each query. The positive and negative distances are calculated:

$$[d_+, d_-] = [\left\| f_v - f_a^+ \right\|_2, \left\| f_v - f_a^- \right\|_2], \tag{5}$$

then these two distances are softmax normalized to $[D_+, D_-]$. The audiovisual correspondence loss function is defined as:

$$\mathcal{L}_{av}(D_+, D_-) = \|D_+\|_2 + \|1 - D_-\|_2 \tag{6}$$

Bottom-up Attention and Top-down Attention Bottom-up attention represents the objectness from a set of basic features like color, size, texture, and shape. Top-down attention is related to prior knowledge and current goals, which in our case are the sounding objects. Their relationship is that bottom-up attention and top-down attention affect each other and are incorporated into the final output visual priority map.

In cognitive science, there is no clear found theory about how this process works. In our implementation, we build the model with the top-down attention map output as the final visual priority map. As we use the derived top-down attention map as the incorporated attention map, we refer to the bottom-up attention map as a guiding restriction to top-down attention. That is to say, the network is supposed to pay more attention to salient objects proposed in bottom-up attention. It accords with our intuition, as when we look for the sounding objects, we tend to look in the objects we find in the scene at first. Therefore, the bottom-up attention map is used as a ground truth-like supervision and a cross entropy loss function can be used to learn the attention correspondence.

In binary classification, the cross entropy loss can be calculated as:

$$\mathcal{L}_{CE} = -\sum_{j=1}^{N} (t_j log(p_j) + (1 - t_j) log(1 - p_j)), \tag{7}$$

where t_j denotes the truth value 0 or 1 and p_j represents the predicted probability of j^{th} sample.

In the context of our top-down attention and bottom-up attention, with A_{bottom} firstly resized to $\lfloor \frac{H}{16} \rfloor \times \lfloor \frac{W}{16} \rfloor$, the attention loss function can be defined as:

$$\begin{aligned}\mathcal{L}_{att}(A_{top},\ A_{bottom}) = -\sum_{i,j} (A_{bottom_{i,j}} log(A_{top_{i,j}}) \\ + (1 - A_{bottom_{i,j}}) log(1 - A_{top_{i,j}})),\end{aligned} \tag{8}$$

where $A_{(i,j)}$ represents the attention value at pixel (i, j) of corresponding attention map. It is a value between 0 and 1. Here $A_{bottom_{i,j}}$ is regarded as a soft label. With an indicator function, it can be defined with a hard label as:

$$\begin{aligned}\mathcal{L}_{att}(A_{top}, A_{bottom}) = -\sum_{i,j} (I_{l_{i,j}} log(A_{top_{i,j}}) \\ + (1 - I_{l_{i,j}}) log(1 - A_{top_{i,j}})),\end{aligned} \tag{9}$$

where $I_{l_{i,j}}$ is a loss indicator function:

$$I_{l_{i,j}} = \begin{cases} 1 & \text{if } A_{bottom_{i,j}} \text{ is larger than threshold } t, \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

However, this is not the way we want to correlate the two forms of attention. This conventional cross entropy function actually encourages top-down attention to attend to all the potential salient regions. We do not want top-down attention to copy bottom-up attention. The bottom-up attention map is supposed to be a guidance during the generation of top-down attention. To this end, we modify the conventional loss function, and the bottom-up attention loss function is defined as:

$$\mathcal{L}_{att}(A_{top}, A_{bottom}) = -\sum_{i,j} (1 - I_{l_{i,j}}) log(1 - A_{top_{i,j}}), \tag{11}$$

where we only keep the negative part of cross entropy function.

Combining the two loss functions mentioned above, the overall unsupervised loss function is defined as:

$$\begin{aligned}\mathcal{L}(f_v, f_a^+, f_a^-, A_{top}, A_{bottom}) = \mathcal{L}_{av}(f_v, f_a^+, f_a^-) \\ + \alpha \mathcal{L}_{att}(A_{top}, A_{bottom}),\end{aligned} \tag{12}$$

where α is a tunable weighting hyperparameter.

	Methods	cIoU@0.5	AUC
Unsupervised	Attention (10k) [29]	43.6	44.9
	Negative Mining (10k) [8]	58.2	52.5
	Attention (144k) [29]	66.0	55.8
	Negative Mining (144k) [8]	69.9	57.3
	Our Bottom-Up Attention (10k)	**73.4**	**57.6**
	Random Bottom-Up Attention (10k)	25.4	35.2
Supervised	CAM (10k) [26]	52.2	49.6
	Sup. Attention (2.5k) [29]	**82.0**	60.7
	Our Bottom-Up Attention (2.5k) with Faster RCNN	80.8	**60.9**

Table 1. Evaluation results of recent sound localization methods on the Flickr-SoundNet dataset.

4. Experiments

4.1. Dataset

The Flickr-SoundNet dataset presented by [5] contains sound and image pairs extracted from more than two million unconstrained videos for cross-modal recognition. [29] sampled a 5k size subset from the Flickr-SoundNet dataset and annotated the sound sources with bounding boxes for supervised learning and qualitative evaluation. This is now the only annotated open dataset for general sounding object localization. For training, We randomly choose 10k samples from the Flickr-SoundNet dataset. For evaluation, random 250 image-audio pairs are chosen from the 5k annotated set.

MUSIC dataset consists of 685 video samples, containing 11 categories of musical instrument. Since this dataset is smaller, we use it for more comprehensive evaluation. We annotated a random subset of 250 samples in the MUSIC dataset in a segmentation way.

4.2. Implementation Details

We implement our framework in PyTorch [25]. Audio signals are sampled at 22050Hz and we take the first 20 seconds (repeat if not long enough). We resize RGB images to 320 x 320. Therefore, the output attention map is 20 x 20. The model is trained by Adam optimizer with betas 0.9 and 0.999. For the unsupervised setting of the bottom-up attention module, we use OpenCV Selective Search [23] to implement it. For simplicity, we set $weight_k$ = $1/K$. For the supervised setting, $weight_k$ is set to the confidence score generated by Faster RCNN, and the attention map values are clipped to the range 0 and 1. We pre-train Faster RCNN with Pascal VOC dataset [11], which contains 20 classes including some common sounding objects like person and animals as well as usually silent objects like chair, table, and plant. If no otherwise specified, region number K is set to 50, loss function threshold t is set to 0.02, object detection confidence threshold τ is set to 0.5 and loss weight α is set to 0.1.

4.3. Results

Quantitative Results. Consensus Intersection over Union (cIoU) [29] is employed as the evaluation metric and 0.5 is set for the cIoU threshold. We compare our methods with recent supervised and unsupervised sound localization methods evaluated on the Flickr-SoundNet dataset. Table 1 shows the evaluation results of different methods. [29] and [8] trained their models with a 10k training set and a 144k training set in an unsupervised way. The supervised attention method in [29] used 2.5k annotated samples. The CAM method [26] leveraged the category labels of images and sounds and established sound-object label alignment. CAM is adopted to measure class-specific correspondence on each spatial grid. Similar to ours, the CAM method [26] does not need human annotations and is trained in an unsupervised way. Given that additional supervision from pre-trained models is provided, we compare it within the supervised field.

In our unsupervised setting, the bottom-up attention module is based on selective search, and in our supervised setting, it is based on a pre-trained Faster RCNN model. The results show that our unsupervised method advances other unsupervised methods by a large margin with only 10k train data needed. It can also be observed that our supervised method can have competitive performance compared with state-of-the-art supervised methods. We repeat that our supervised method does not need extra human annotations.

For comparison, we also conduct experiments with a random generated bottom-up attention map. The results show that this random attention map decreases performance. It confirms that our bottom-up attention provides meaningful guidance for top-down attention.

We further evaluate our method pretrained on the SoundNet-Flickr dataset on MUSIC dataset. Compared with unsupervised baseline with audiovisual attention [28], our bottom-up attention method improved the accuracy by +17%, +9%, and+3% with IoU thresholds at 0.2, 0.3 and 0.5 respectively.

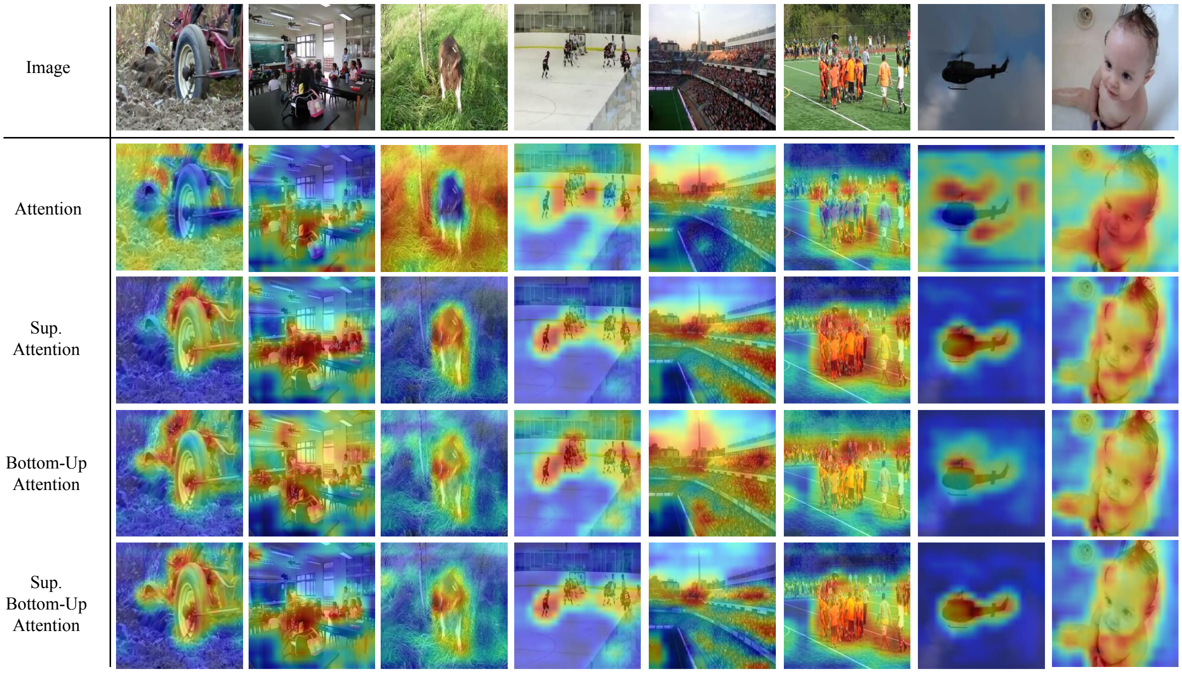

Figure 5. Qualitative sound localization results on the Flickr-SoundNet dataset using unsupervised and supervised attention models [29] and our models (the last two rows of pictures).

Qualitative Results. Figure 5 visualizes the localization results of the image-sound pairs from the Flickr-SoundNet dataset [5] using our unsupervised and supervised bottom-up attention methods and the unsupervised and supervised attention models present in [29]. It shows that our unsupervised method achieves comparable performance to the supervised methods.

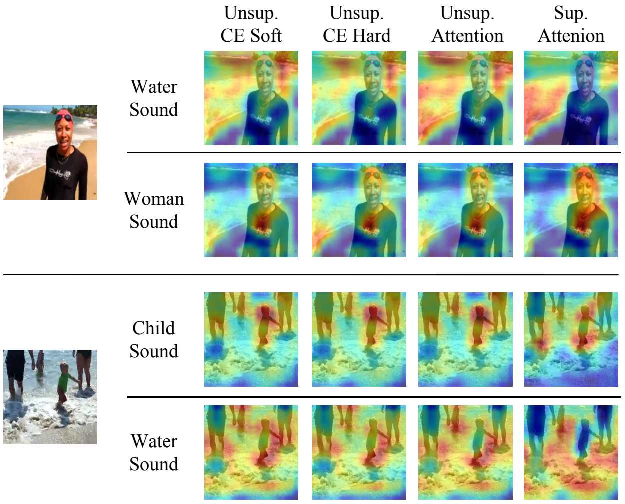

Figure 6. **Qualitative sound discrimination evaluation in different learning settings.** All settings are trained with the bottom-up attention mechanism. Results of different loss functions are shown.

To present our model's sound discrimination ability, we also visualize the responses of the network to different sounds while keeping the frame the same. The results are shown in Figure 6. It shows that our network can distinguish different sounds instead of simply locating the salient objects in the scene. For example, given an image of a woman walking along the beach, hearing the sound of a woman talking, the region of the woman should be paid more attention. In contrast, hearing the sound of the sea, the sea surface should be attended to. For better comparison of different loss functions in Section 3.3, we also present the results under the same setting while training with the soft label cross entropy loss function defined in Equation 9 and hard label cross entropy loss function defined in Equation 8. It confirms that our modified loss function effectively improves the discrimination ability. The qualitative comparison will be discussed later.

The evaluation results suggest that our top audiovisual attention is generated with the guidance of bottom-up attention but is not restricted to the latter. The audiovisual attention regions are not determined by object region proposals.

4.4. Ablation Experiments

Unsupervised Setting. The impact of the number of kept region proposals K and loss function threshold t is summarized in Table 2. We implement it with $weight_k = 1/K$ for simplicity. The results suggest that a proper number of region proposals is necessary. Too few regions cannot de-

K	t	cIoU@0.5	AUC
50	0.02	72.4	56.8
50	0.04	68.4	55.5
50	0.06	65.8	54.8
100	0.01	71.2	57.2
100	0.02	**73.4**	**57.6**
100	0.04	47.2	26.4
200	0.005	47.6	24.3
200	0.01	46.8	27.0
200	0.02	45.2	25.7

Table 2. **Ablation experiments in the unsupervised setting.** The impact of the number of kept region proposals K and loss function threshold t is reported.

τ	t	cIoU@0.5	AUC
0.3	0.3	79.2	60.5
0.3	0.5	76.8	59.6
0.3	0.7	78.0	59.1
0.5	0.5	**80.8**	**60.9**
0.5	0.7	80.0	60.3
0.7	0.7	79.2	60.1

Table 3. **Ablation experiments in the supervised setting.** The impact of the Faster RCNN detection confidence threshold τ and loss function threshold t is reported.

scribe the objectness of the whole image well, while too many will import too much noise. We visualize the bottom-up attention maps with K set to 50, 100, and 200 respectively in Figure 7. It shows that with K increasing, bottom-up attention tends to cover more salient areas. However, when too much noise is imported, the guidance effect of bottom-up attention decreases.

Supervised Setting. For each sample, our pre-trained Faster RCNN [28] generates 6000 bounding boxes, its corresponding predicted category, and a confidence score. As described in Section 3.1, we keep the bounding boxes whose confidence scores are larger than a threshold τ, while we ignore the predicted class here. Table 3 shows the results of the ablation experiments on this confidence threshold τ and loss function threshold t. The results suggest that although τ is insensitive in general due to usually high confidence scores, a proper threshold provides a better description of the objectness.

Loss Functions. To analyze the performance of our present bottom-up attention loss function, we conduct experiments based on different loss functions. The results are shown in Table 4. Our attention loss function is defined in Equation 11, and cross entropy loss functions with soft label and hard label are defined in Equation 8 and Equation 9 respectively.

Loss	α	cIoU@0.5	AUC
CE with Soft Label	0.1	69.2	55.3
	0.05	66	55.4
CE with Hard Label	0.1	66.4	55.2
	0.05	66	55.5
Ours	0.1	**71.2**	**56.7**
	0.05	69.2	56

Table 4. Results of different loss functions.

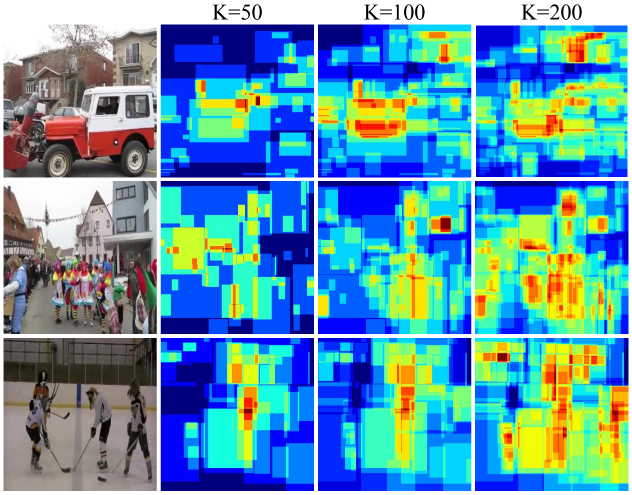

Figure 7. Bottom-up attention maps with K set to 50, 100, and 200 respectively.

The results demonstrate that our modified loss function improves the localization ability.

5. Conclusions

In this paper, we focus on the task of locating sounding objects in unconstrained visual scenes. We present an unsupervised method based on bottom-up attention and top-down attention. Top-down attention captures the audiovisual correspondence under the guidance of bottom-up attention. We also present a novel bottom-up attention loss to learn the correlation between the two forms of attention. Our proposed unsupervised method advances other unsupervised methods by a large margin on the public sound localization dataset. Our method implemented in the supervised setting also achieves competitive performance to the latest supervised methods.

Acknowledgements

This work was supported by NSFC (61906119, U19B2035), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2189–2202, 2012.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

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

[4] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014.

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[7] Joao Carreira and Cristian Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1312–1328, 2011.

[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 (CVPR)*, pages 16867–16876, June 2021.

[9] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer, 2013.

[10] Ian Endres and Derek Hoiem. Category independent object proposals. In *European Conference on Computer Vision*, pages 575–588. Springer, 2010.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[12] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

[13] 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.

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

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

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

[19] 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. *arXiv preprint arXiv:2010.05466*, 2020.

[20] Di Hu, Zheng Wang, Haoyi Xiong, Dong Wang, Feiping Nie, and Dejing Dou. Curriculum audiovisual learning. *arXiv preprint arXiv:2001.09414*, 2020.

[21] Fumi Katsuki and Christos Constantinidis. Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 20(5):509–521, 2014.

[22] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[23] OpenCV. Open source computer vision library, 2015.

[24] 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.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[26] Rui Qian, D Hu, H Dinkel, M Wu, N Xu, and W Lin. A two-stage framework for multiple sound-source localization. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, volume 19, page 20, 2020.

[27] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[29] 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.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[32] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2745–2754, June 2021.

[33] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[34] 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.

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[36] 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.