Localizing Visual Sounds the Easy Way

Shentong Mo[1] and Pedro Morgado[1,2(✉)]

[1] Carnegie Mellon University, Pittsburgh, USA
[2] University of Wisconsin-Madison, Madison, USA
pmorgado@wisc.edu

Abstract. Unsupervised audio-visual source localization aims at localizing visible sound sources in a video without relying on ground-truth localization for training. Previous works often seek high audio-visual similarities for likely positive (sounding) regions and low similarities for likely negative regions. However, accurately distinguishing between sounding and non-sounding regions is challenging without manual annotations. In this work, we propose a simple yet effective approach for Easy Visual Sound Localization, namely EZ-VSL, without relying on the construction of positive and/or negative regions during training. Instead, we align audio and visual spaces by seeking audio-visual representations that are aligned in, at least, one location of the associated image, while not matching other images, at any location. We also introduce a novel object guided localization scheme at inference time for improved precision. Our simple and effective framework achieves state-of-the-art performance on two popular benchmarks, Flickr SoundNet and VGG-Sound Source. In particular, we improve the CIoU on Flickr SoundNet from 76.80% to 83.94%, and on VGG-Sound Source from 34.60% to 38.85%. Code and pretrained models are available at https://github.com/stoneMo/EZ-VSL.

1 Introduction

When we hear a baby crying, we can localize the sound by finding the baby in the room. This ability of visual sound source localization is possible due to the tight association between visual and auditory signals in the natural world. In this work, we aim to leverage this natural and freely available audio-visual association to localize sound sources present in a video in an unsupervised manner, i.e. without relying on manual annotations for sounding source locations.

Unsupervised visual localization of sound sources has attracted much attention in recent years [2,6,30]. To tackle this problem, recent approaches [2,6,17, 28,31] rely on direct audio-visual similarity in a learned latent space for localization. These audio-visual similarities are used to construct likely sounding and non-sounding regions in the image, and the models are learned by requiring the audio representation to match visual representations pooled from likely sounding regions while being dissimilar from those of different images [2,17,28], and/or from non-sounding regions [6,31]. While these approaches have been shown to

yield state-of-the-art performance in unsupervised visual sound localization, we identify two major limitations.

First, the training objective presents a paradox. On one hand, accurate regions of sounding objects are required in order to encourage audio representations to match the visual representations of the regions where the source is located. On the other hand, since localization maps are obtained through audio-visual similarities, accurate representations are required in order to identify the regions containing the sounding objects. This paradox results in a complex training objective that is likely to contain many sub-optimal local minima, as the model is required to bootstrap from its own localization ability.

Second, by solely relying on audio-visual similarity for localization, prior work ignores the visual prior of likely audio sources. For example, even without access to the audio signal, we know that most regions of an image, depicting for example the floor, the sky, a table, or a wall, are unlikely to depict sources of sound.

(a) Flick SoundNet   (b) VGG Sound Sources
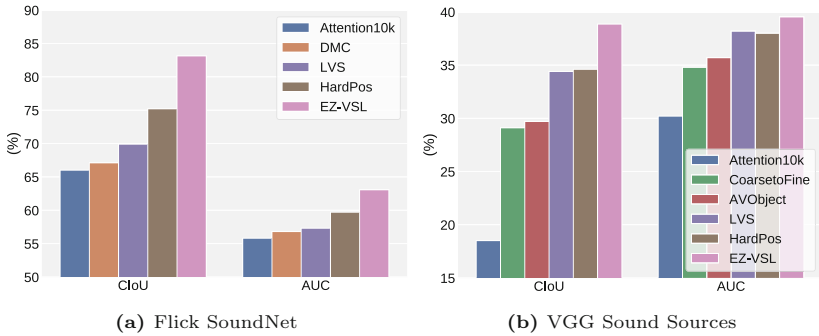
Fig. 1. Comparison of EZ-VSL with state-of-the-art methods on Flickr SoundNet [17] (a) and VGG-SS [6]. All methods in (a) are trained on Flickr 144k, and those in (b) on VGG Sound 144k.

To address these challenges, we propose a simple yet effective approach for easy visual sound localization, namely EZ-VSL. Instead of relying on explicit maps for sounding and non-sounding regions, we treat audio-visual correspondence learning as a multiple instance learning problem. In other words, we propose a training loss that encourages the audio signal to be associated with, at least, one location in the corresponding image, while not being associated with any location from other images. Then, we introduce a novel object-guided localization scheme at inference time that combines the audio-visual similarity map with an object localization map from a lightweight pre-trained visual model, which biases sound source localization predictions towards the objects in the scene.

We evaluate our EZ-VSL on two popular benchmarks, Flickr SoundNet [17] and VGG-Sound Source [6]. Extensive experiments show the superiority of our

approach for unsupervised sound source visual localization. We also conduct comprehensive ablation studies to demonstrate the effectiveness of each component. Surprisingly, we found that the object prior alone, which does not even leverage the audio for localization, already surpasses all prior work on both Flickr and VGG-Sound benchmarks. We also demonstrate the superiority of the proposed multiple instance learning objective for audio-visual matching compared to prior approaches that rely on careful constructions of positive (sounding) and negative (non-sounding) regions for training. Finally, we show that the visual object prior and audio-visual similarity maps can be further combined into more accurate predictions, surpassing the current state-of-the-art method by large margins on both Flickr SoundNet and VGG Sound Sources. These results are highlighted in Fig. 1.

Overall, the main contributions of this work can be summarized as follows:

✦ We present a simple yet effective multiple instance learning framework for unsupervised sound source visual localization, which we call EZ-VSL.
✦ We propose a novel object-guided localization scheme that favors object regions, which are more likely to contain sound sources.
✦ Our EZ-VSL successfully achieves state-of-the-art performance on two popular benchmarks, Flickr SoundNet and VGG-Sound Source.

2   Related Work

Audio-Visual Joint Learning. Several works [3,4,21–25,27,33,34] have been proposed in recent year on audio-visual self-supervised learning to learn bimodal representations from each other. SoundNet [4] applies a visual teacher network to extract audio representations from untrimmed videos. The audio-visual correspondence task [3] is introduced to learn both visual and audio representations in an unsupervised way. Audio-visual synchronization objectives are also explored for several tasks, such as speech recognition [1,32], audio-visual navigation [5], visual sound source separation, and localization [10,12,14,30,35,36].

Besides these works, several methods adopt a weakly-supervised scheme to solve audio-visual problems. For example, UntrimmedNet [34] uses a classification module and a selection module for Multiple Instance Learning (MIL) to perform audio-visual action localization. [33] also proposes in a hybrid attention network for audio-visual video parsing. In this work, however, we focus on the sound source localization problem by learning audio-visual representations jointly from unlabelled videos.

Audio-Visual Source Localization. Audio-Visual Source Localization aims at localizing sound sources by learning the co-occurrence of audio and visual features in a video. Early works [9,16,19] use shallow probabilistic models or canonical correlation analysis to solve this problem. With the introduction of deep neural networks, some approaches [17,26] were proposed to learn the audio-visual correspondence via a dual-stream network and a contrastive loss. For instance, DMC [17] adopts synchronous sets of clustering with respect to each

modality for capturing audio-visual correspondences. Multisensory features [26] are used to jointly learn visual and audio representations of a video through the temporal alignment. Other methods [11,13,29,35,36] leverage the audio-visual source separation as the target to achieve visual sound localization. Most of these methods learn from global audio-visual correspondences. Although they show qualitatively that the model is capable of localization, their localization ability is not competitive to models that learn from localized correspondences.

Beyond the work discussed above, several relevant works have targeted the visual source localization problem directly. Attention10k [30] developed an attention mechanism and a two-stream architecture with each modality to localize sound sources in an image. Qian *et al.* [28] proposed a two-stage framework to learn audio and visual representations with the cross-modal feature alignment in a coarse-to-fine way. Afouras *et al.* [2] introduced an attention-based model with the optical flow to localize and group sound sources in a video. More recently, LVS [6] added a hard sample mining mechanism to contrastive loss with a differentiable threshold on the audio-visual correspondence map. Finally, Hard-Pos [31] leveraged hard positives in contrastive learning for learning semantically matched audio-visual information from negative pairs. Different from these baselines, we show that it is possible (and even preferable) to learn from a simplified multiple-instance contrastive learning objective. We also propose a novel object guided localization scheme to boost the visual localization performance of sound sources.
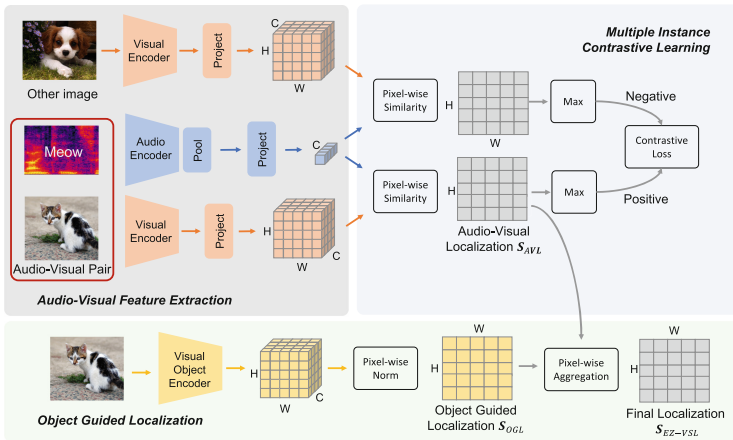
Fig. 2. Illustration of the proposed method. The audio-visual feature extractor computes global audio and localization visual features. Audio-visual alignment is learned using a multiple instance contrastive learning objective. At inference time, we use another visual encoder pre-trained on object recognition to compute object localization maps, which are combined with audio-visual localization maps for the final prediction.

3   Method

Given a video containing sound sources, our goal is to localize the sounding objects within it without using manual annotations of their locations for training. We propose a simple yet effective way for unsupervised sound source visual localization, which we denote EZ-VSL.

3.1   Overview

Let $\mathcal{D} = \{(v_i, a_i) : i = 1, \ldots, N\}$ be a dataset of paired audio a_i and visual data v_i, where the sources of the sound audible in a_i are assumed to be depicted in v_i. Following previous work [6,17], we first encode the audio and visual signals using a two stream neural net encoder, denoted as $f_a(\cdot)$ and $f_v(\cdot)$ for the audio and images, respectively. The audio encoder extracts global audio representations $\mathbf{a}_i = f_a(a_i)$ and the visual encoder computes localized representations $\mathbf{v}_i^{xy} = f_v(v_i^{xy})$ for each (x, y) location. As shown in Fig. 2, audio and visual features are then mapped into a shared latent space, where the similarity between audio-visual representations can be computed for all locations. The audio-visual models are then trained to minimize a **cross-modal multiple-instance contrastive learning** loss, that encourages audio representation to be aligned with the associated visual representations at least at one location. By optimizing this loss, audio and visual signals are matched in the shared latent space, which can then be used for localization. At inference time, we combine the learned audio-visual similarities with **object guided localization**. We accomplish this using a visual model pre-trained for object recognition. It should be noted that models pre-trained on ImageNet are already used to initialize the visual encoder for VSL [2,6,17,28,30,31]. We use the same model to extract regions of the image that are likely to contain objects (regardless of whether they are producing the sound or not). The object maps are then integrated with audio-visual similarities to enhance localization accuracy.

We now elaborate on the two main components of our work: the multiple instance contrastive learning objective, and the object guided localization.

3.2   Audio-Visual Matching by Multiple-Instance Contrastive Learning

Aligning audio and localized visual representations poses two main challenges. First, the output of the audio and visual encoders are not necessarily compatible. Second, most locations in the image do not depict the sound source, and so the representations at these locations should not be aligned with the audio.

The first challenge can be easily addressed by projecting both audio and visual representations into a shared feature space

$$\hat{\mathbf{v}}_i^{xy} = \mathbf{U}_v \mathbf{v}_i + \mathbf{b}_v \quad \forall x, y, i \qquad \text{and} \qquad \hat{\mathbf{a}}_i = \mathbf{U}_a \mathbf{a}_i + \mathbf{b}_a \quad \forall i \qquad (1)$$

where $\mathbf{U}_v$ and $\mathbf{U}_a$ are projection matrices, and $\mathbf{b}_v$ and $\mathbf{b}_a$ bias terms.

The second challenge requires to selectively match the audio representations to the associated visual regions depicting the sound sources. Prior work [2,6,17, 28,30,31] explicitly computes an attention map for the likely sounding regions by bootstrapping from current audio-visual similarities. The audio representations are then required to match these sounding regions [17,28,30], and in some cases to not match non-sounding regions from the same image [6,31]. As discussed above, this leads to a paradox where accurate localization is required to learn accurate audio-visual representations, which is required for localization in the first place.

To simplify this framework, we propose to optimize a multiple instance contrastive learning loss. Each bag of visual features V (or bag of instances) spans all locations within an image

$$V_i = \{\hat{\mathbf{v}}_i^{xy} : \forall x, y\} \quad \forall i \in \mathcal{D} \tag{2}$$

Audio representations $\mathbf{a}_i$ are then required to be similar to at least one instance in the corresponding positive bag V_i, while being dissimilar from all locations in all negative bags $V_j \; \forall j \neq i$. Specifically, we seek to maximize the alignment between the audio and the most similar positive visual instance, through the following loss function

$$\mathcal{L}_{a\to v} = -\log \frac{\exp\left(\frac{1}{\tau} \max_{\hat{\mathbf{v}}\in V_i} \mathtt{sim}(\hat{\mathbf{a}}_i, \hat{\mathbf{v}})\right)}{\sum_k \exp\left(\frac{1}{\tau} \max_{\hat{\mathbf{v}}\in V_k} \mathtt{sim}(\hat{\mathbf{a}}_i, \hat{\mathbf{v}})\right)} \tag{3}$$

where $\mathtt{sim}(\hat{\mathbf{v}}, \hat{\mathbf{a}}) = \hat{\mathbf{v}}^T \hat{\mathbf{a}}/(\|\hat{\mathbf{v}}\|\|\hat{\mathbf{a}}\|)$ is the cosine similarity, and τ a temperature hyper-parameter. Negative bags are obtained from other samples in the same mini-batch. To train our models, we use a symmetric version of (3) by defining

$$\mathcal{L}_{v\to a} = -\log \frac{\exp\left(\frac{1}{\tau} \max_{\hat{\mathbf{v}}\in V_i} \mathtt{sim}(\hat{\mathbf{v}}, \hat{\mathbf{a}}_i)\right)}{\sum_k \exp\left(\frac{1}{\tau} \max_{\hat{\mathbf{v}}\in V_i} \mathtt{sim}(\hat{\mathbf{v}}, \hat{\mathbf{a}}_k)\right)}, \tag{4}$$

and optimizing the symmetric loss

$$\mathcal{L} = \mathcal{L}_{a\to v} + \mathcal{L}_{v\to a}. \tag{5}$$

During inference, the audio-visual localization map is computed as

$$\mathbf{S}_{xy}^{AVL} = \mathtt{sim}(\hat{\mathbf{v}}_{xy}, \hat{\mathbf{a}}) \quad \forall x \in [1, W], y \in [1, H]. \tag{6}$$

3.3   Object-Guided Localization

At inference time, we propose a novel object-guided scheme for enhanced localization. The input image is fed to a convolutional model f_{obj} pre-trained on ImageNet [8] without global pooling or the classification head, yielding a feature map $\mathbf{v}' = f_{obj}(v) \in \mathbb{R}^{C \times H \times W}$. This model has the same architecture than the visual encoder used for audio-visual localization and is initialized with the same ImageNet pre-trained weights, but unlike the former, this model is never trained

for audio-visual similarity. Hence, the feature map $\mathbf{v}'$ contains zero information about the accompanying audio. Instead, it can be used to define a localization prior that favors the objects in the scene, regardless of whether these objects are the sources of the sound or not. We then experimented with two possible solutions to extract object-centric localization maps without any additional training. The first obtains a 1000-way object class posterior $P(o|\mathbf{v}'_{xy})$ by applying an ImageNet pretrained classifier to each (x,y) location of $\mathbf{v}'$. We then define the object localization prior as

$$\mathbf{S}^{CLS}_{xy} = \max_o P(o|\mathbf{v}'_{xy}). \tag{7}$$

The second approach, perhaps less intuitive but more effective, relies on the fact that f_{obj} was trained on an object-centric dataset, and thus produces stronger activations when evaluated on images of objects. With this intuition in mind, we alternatively define the object localization prior as

$$\mathbf{S}^{L1}_{xy} = \|\mathbf{v}'_{xy}\|_1. \tag{8}$$

Note that in both cases, the object prior solely relies on a model f_{obj} pre-trained on ImageNet. We conduct no further training of f_{obj}.

The audio-visual localization and object-centric maps are then linearly aggregated into a final localization map $\mathbf{S}^{EZVSL}_{xy}$ of the form

$$\mathbf{S}^{EZVSL}_{xy} = \alpha\mathbf{S}^{AVL}_{xy} + (1-\alpha)\mathbf{S}^{OBJ}_{xy} \quad \forall x,y, \tag{9}$$

where $\mathbf{S}^{AVL}_{xy}$ is the audio-visual similarity of map of (6), $\mathbf{S}^{OBJ}_{xy}$ is the object localization map (i.e., $\mathbf{S}^{CLS}_{xy}$ in (7) or $\mathbf{S}^{L1}_{xy}$ in (8)), and α is balancing term that weights the contribution of the object prior and the audio-visual similarity terms. In practice, since the two maps $\mathbf{S}^{AVL}_{xy}$ and $\mathbf{S}^{OBJ}_{xy}$ can have widely different ranges of scores, we normalize them into a $[0,1]$ range before aggregation, i.e., $\mathbf{S}_{xy} = \frac{\mathbf{S}_{xy} - \min_{xy}\mathbf{S}_{xy}}{\max_{xy}\mathbf{S}_{xy} - \min_{xy}\mathbf{S}_{xy}}$.

4   Experiments

We evaluated EZ-VSL on unsupervised visual sound source localization. Following accepted practices [6,28,30], we used the Flickr SoundNet dataset [4] and the recently proposed VGG-Sound dataset [7], and report the same evaluation metrics as in [6,28,30]. Namely, we measure the average precision at a Consensus Intersection over Union threshold of 0.5, a metric often simply denoted as CIoU. We also measure the Area Under Curve (AUC).

4.1   Experimental Setup

Datasets. Flickr SoundNet includes 2 million unconstrained videos from Flickr. From each video clip, a single image frame is extracted together with 20 s of audio centered around it, to form the corresponding audio-visual pairs used for

unsupervised learning. We also conduct experiments on **VGG-Sound** composed of 200k video clips from 309 sound categories. Similar to the Flickr dataset, the video is represented by a single frame as well as its audio. To enable direct comparisons with existing work [2,6,28,30], we trained our models using subsets of either 10k or 144k image-audio pairs.

Localization performance is measured on two datasets, the Flickr SoundNet test set [30] and the more challenging VGG-Sound Sources test set [6]. The former includes only 250 image-audio pairs for which the location of the sound source has been manually annotated. The latter contains annotations for 5000 instances spanning 220 sounding objects categories.

Table 1. Localization performance on the Flickr SoundNet testset.

Method	CIoU (%)	AUC (%)	Method	CIoU (%)	AUC (%)
Attention10k [30]	43.60	44.90	Attention10k [30]	66.00	55.80
CoarsetoFine [28]	52.20	49.60	DMC [17]	67.10	56.80
AVObject [2]	54.60	50.40	LVS [6]	69.90	57.30
LVS [6]	58.20	52.50	HardPos [31]	75.20	59.70
EZ-VSL (ours)	**81.93**	**62.58**	EZ-VSL (ours)	**83.13**	**63.06**

(a) Training set: Flickr 10k	**(b)** Training set: Flickr 144k

Audio and Visual Pre-processing. The input to the visual encoder $f_v(\cdot)$ are images of resolution 224×224. During training, images are first resized to 246 along the shortest edge, and random cropping together with random horizontal flipping is applied for data augmentation. At test time, images directly resized into a 224×224 resolution without cropping.

The audio encoder $f_a(\cdot)$ takes the log spectrograms extracted from 3 s of audio extracted at a sample rate of 11025 Hz. The underlying STFT are computed using approximately 50 ms windows with a hop size of 25 ms, resulting in an input tensor of size 257×300 (257 frequency bands over 300 timesteps). No data augmentations are applied during train or test time.

Audio and Visual Models. Both the visual and audio encoders are implemented using the lightweight ResNet18 [15] as the backbone. Following prior work [6,17,28], we initialized the visual model using weights pre-train on ImageNet [8]. Unless otherwise specified, the audio and visual representations are projected into a shared space of dimension 512.

The model is trained with a batch size of 128 on 2 GPUs. For efficiency, we only use negatives from the local batch, i.e. we did not gather negatives from all GPUs. This results in a negative set of 63 samples for the contrastive learning objective of (3). The model is trained using the Adam optimizer [20] with a learning rate of $1e - 4$, and default hyper-parameters $\beta_1 = 0.9, \beta_2 = 0.999$. On large datasets (144k or the full VGG-Sound database), the model is trained for 20 epochs. On smaller (10k) datasets, the model is trained for 100 epochs.

Table 2. Localization performance on Flickr SoundNet and VGG-SS after training on VGG-Sound 144k.

Training set	Method	Flickr-SoundNet		VGG-SS	
		CIoU (%)	AUC (%)	CIoU (%)	AUC (%)
VGG-Sound 144k	Attention10k [30]	66.00	55.80	18.50	30.20
	CoarsetoFine [28]	-	-	29.10	34.80
	AVObject [2]	-	-	29.70	35.70
	LVS [6]	73.50	59.00	34.40	38.20
	HardPos [31]	76.80	59.20	34.60	38.00
	EZ-VSL (ours)	**83.94**	**63.60**	**38.85**	**39.54**

4.2   Comparison to Prior Work

In this work, we propose a simple yet highly effective training framework for visual sound source localization. To demonstrate the effectiveness of our approach, EZ-VSL, we start by drawing direct comparisons to previous works [2, 6,28,30] on two popular benchmarks: Flickr SoundNet [17] and VGG-SS [6]. Results are reported in Tables 1 and 2 for models trained on Flickr SoundNet and VGG-SS, respectively.

As can be seen, EZ-VSL outperforms prior work by large margins, establishing new state-of-the-art results in all settings. On the Flickr test set, we observe performance gains of 23.73% CIoU and 10.08% AUC when models are trained on Flickr 10k, by 7.93% CIoU and 3.36% AUC when trained on Flickr 144k, and by 7.14% CIoU and 4.4% AUC when trained on VGG-Sound 144k. Significant gains can also be observed on the more challenging VGG-Sound Sources test set, with EZ-VSL outperforming prior work by 4.25% CIoU and 1.34% AUC.

We highlight that these gains are obtained with a significantly simplified training objective. For example, Attention10K [30] relies on the construction of positive (sounding) regions for its visual attention mechanism, and both LVS [7] and HardPos [31] require not only the construction of likely positive (sounding) regions but also negative (non-sounding) regions. This highlights the importance of a well-designed training framework that avoids imposing complex region-specific constraints. Also, note that our method combines both the novel multiple instance contrastive learning loss used for training and the novel object-centric localization procedure used during inference. The effect of these individual components will be studied below.

4.3   Open Set Audio-Visual Localization

To assess generalization, we evaluated the ability of EZ-VSL to generalize beyond the categories of sound sources heard during self-supervised training. Following previous work [6], we randomly sampled 110 categories from VGG-Sound for training. We then evaluate our model on test samples from these heard categories, as well as on samples from another 110 unheard categories. Since unseen

categories can be semantically related to the seen ones, we expect that good representations to generalize to unseen categories as well. The results are shown in Table 3. As can be seen, our approach outperforms LVS [6] by a significant margin on both heard and unheard categories. In fact, unlike LVS, the performance of our EZ-VSL model did not suffer by the presence of unheard sound categories, achieving even slightly better performance on unheard classes than on heard classes. This provides evidence for the stronger generalization ability of EZ-VSL in an open set setting.

Table 3. Comparison results on VGG-SS for open set audio-visual localization trained on 70k data with heard 110 classes.

Test class	Method	CIoU (%)	AUC (%)
Heard 110	LVS [6]	28.90	36.20
	EZ-VSL	**37.25**	**38.97**
Unheard 110	LVS [6]	26.30	34.70
	EZ-VSL	**39.57**	**39.60**

4.4   Cross Dataset Generalization

To further evaluate generalization, we tested models across datasets. Specifically, we tested the model trained on VGG-Sound on Flickr SoundNet, and test the Flickr trained model on the VGG-SS test set. As can be seen in Table 4, our approach outperforms the best previous method [6] when testing across datasets.

Table 4. Cross dataset generalization results of Flickr SoundNet and VGG-SS trained on various training sets, including VGG-Sound 10k, 144k, Full and Flickr 10k, 144k.

Test set	Training set	Method	CIoU (%)	AUC (%)
Flickr SoundNet	VGG-Sound 10k	LVS [6]	61.80	53.60
		EZ-VSL	**78.71**	**61.56**
	VGG-Sound 144k	LVS [6]	71.90	58.20
		EZ-VSL	**84.34**	**63.77**
	VGG-Sound Full	LVS [6]	73.59	59.00
		EZ-VSL	**83.94**	**63.60**
VGG-SS	Flickr 10k	LVS [6]	18.71	30.29
		EZ-VSL	**35.54**	**38.18**
	Flickr 144k	LVS [6]	26.95	34.30
		EZ-VSL	**38.62**	**39.20**

4.5   Experimental Analysis

We conducted extensive ablation studies to explore the benefits of the two main components of our approach: multiple instance contrastive learning (MICL) and object-guided localization (OGL). We also conducted several parametric studies to assess the impact of hyper-parameters such as the size of shared audio-visual latent space, the audio-visual fusion strategy, or the balancing coefficient α used for OGL. All experiments were trained on the VGG-Sounds full training set and evaluated on Flickr-SoundNet and VGG-Sound Source (VGG-SS) test sets.

Disentangling the Benefits of MICL and OGL. We ablated the use of MICL and OGL to verify their effectiveness. Models evaluated without MICL only use the object guided localization maps extracted from the pre-trained ResNet-18, without any further training. Models evaluated without OGL only use the audio-visual localization (AVL) maps learned using MICL. We further evaluate two strategies for OGL, namely, classification based OGL (CLS-OGL) described in (7) and activation based OGL (L1-OGL) described in (8).

Results are shown in Table 5. Comparing the performance of each component in isolation (first three rows of Table 5) to those in Table 2, we highlight that both AVL and L1-OGL already surpass prior state-of-the-art (LVS [6]). The strong performance of L1-OGL is especially noteworthy, as it does not even use the audio. We attribute this result to two reasons. First, object regions are more likely to depict sound sources. Second, the majority of test samples in both Flickr and VGG-SS only contain a single sounding object in the scene. This is more prevalent in Flickr but is still true for VGG-SS. As a result, the object prior already provides strong localization results, outperforming all prior work. We nevertheless improve over OGL, by combining it with audio-visual localization.

Table 5. Ablation study on the impact of audio-visual localization (AVL) maps and two object-guided localization strategies (CLS and L1 prior) during inference.

AVL	L1-OGL	CLS-OGL	Flickr SoundNet		VGG-SS	
			CIoU (%)	AUC (%)	CIoU (%)	AUC (%)
✓			78.31	61.74	35.96	38.20
	✓		78.31	61.17	36.77	38.69
		✓	75.10	58.18	35.13	38.08
✓		✓	81.93	62.50	38.58	39.59
✓	✓		**83.94**	**63.60**	**39.34**	**39.78**

Among the two OGL strategies, L1-OGL was the most effective, and thus used as the default strategy for EZ-VSL. We also evaluated the localization performance for various values of the balancing coefficient α between AVL and L1-OGL localization maps. The results in Fig. 3 show that both OGL and AVL components are important for accurate localization, as $\alpha = 0$ or $\alpha = 1$ yields the

worse performance. The optimal value of α for Flickr was 0.4 and for VGG-SS was 0.5. $\alpha = 0.4$ was used as the default for all experiments in this paper.

Dimensionality of Shared Audio-Visual Latent Space. The impact of the latent space dimensionality is shown in Fig. 4. The models were trained on VGG-Sound with latent space of size 32, 64, 128, 256, 512, 1024, 2048, 4096, and tested on Flickr SoundNet and VGG-SS. Figure 4 shows that significantly reducing or increasing the unimodal feature dimensionality (512) can have a negative impact on performance.

Audio-Visual Matching Strategy During Training. The proposed EZ-VSL method uses a max pooling strategy for measuring the similarity between the global audio feature A and the bag of localized visual features $V = \{V_{xy} : \forall x, y\}$, i.e., using $\mathrm{MaxPool}_{xy}(\mathtt{sim}(V_{xy}, A))$. We validate this strategy by comparing two alternatives. First, average pooling is a popular strategy for gathering responses across instances in a bag [18]. We follow this approach and train a model that seeks to match the global audio feature to the visual features at *all* locations, i.e., using $\mathtt{sim}(\mathrm{AvgPool}_{xy}(V_{xy}), A)$. Second, prior work on audio-visual representation learning [3,21,25,27] learn by matching global features. We also tested this class of methods by training a model that pools the visual features before matching to the audio, i.e., using $\mathtt{sim}(\mathrm{MaxPool}_{xy}(V_{xy}), A)$.

The localization performance of all three strategies are reported in Table 6. Since only audio-visual localization maps are impacted by the different training strategies, we set $\alpha = 1$ in this experiment to ignore object-guided localization maps. As can be seen, the two alternative strategies failed to localize sounding objects accurately. On one hand, matching global features lacks the ability to learn localized representations. On the other hand, forcing the audio to match the image at all locations is also inherently problematic, since most regions do not contain a sounding object. The proposed approach achieves significantly better localization performance. However, it assumes that there is at least one sounding object visible in the image. While this is generally true in both VGG-Sound and Flickr SoundNet training sets, further experiments on datasets with non-visible sound sources would be required to assess the robustness of EZ-VSL to this more challenging training scenario.
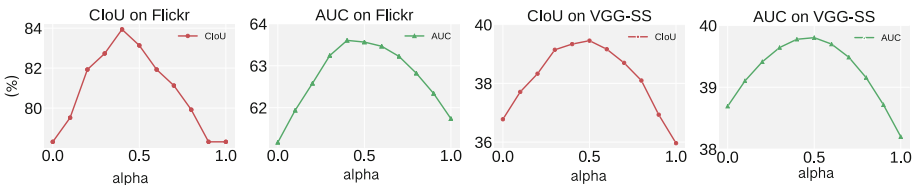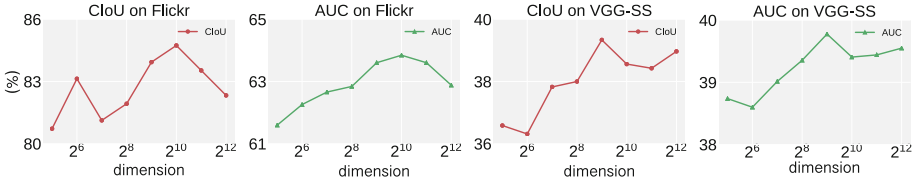

Fig. 3. Impact of α (trade-off between AVL and OGL) on EZ-VSL performance.

Multiple Sound Source Localization. Since complex scenes are known to be more challenging for localization methods, the VGG-SS dataset provide a further

Fig. 4. Impact of the output dimensionality on EZ-VSL performance.

Table 6. Impact of different audio-visual matching strategies during training on audio-visual localization performance. V_{xy} denotes the visual embedding at location (x, y), and A the global audio embedding. Only the audio-visual localization maps are evaluated in this experiment, without being merged with object-guided localization maps.

AV matching strategy	Flickr SoundNet		VGG-SS	
	CIoU (%)	AUC (%)	CIoU (%)	AUC (%)
$\text{sim}(\text{MaxPool}_{xy}(V_{xy}), A)$	49.40	48.97	12.72	27.10
$\text{AvgPool}_{xy}(\text{sim}(V_{xy}, A))$	33.33	37.56	6.03	19.44
$\text{MaxPool}_{xy}(\text{sim}(V_{xy}, A))$	**78.31**	**61.74**	**35.96**	**38.20**

breakdown of test samples per the number of objects. As shown in Fig. 5, similar to prior work, the performance of EZ-VSL does degrade as the scene becomes more complex. However, EZ-VSL consistently outperform prior work regardless of the number of objects.

4.6   Qualitative Results

To better understand the capabilities of the learned model, we show in Fig. 6 sound localization predictions of an EZ-VSL model trained on the VGG-SS 144k dataset. As can be seen, the model is capable of accurately localizing a wide variety of sound sources, showing high overlap with the ground-truth bounding boxes. For example, in row 2, column 4, the model was able to identify that the sound sources are the musical instruments and not the people playing them, or that the sound source in row 3 column 2 is the dog (and not the man). We also show failure cases in Fig. 7. We notice that the learned model often has trouble predicting tight localization maps for small objects, or localizing the sound of crowds, such as in stadiums.

Finally, we compare the final localization map with the object-guided map and the audio-visual similarity map in Fig. 8. These results demonstrate the effectiveness of combining object-guided and audio-visual localization in visual sound localization.
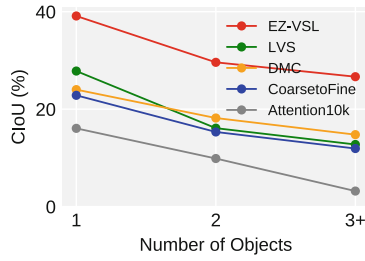
Fig. 5. Localization performance vs number of objects in the scene. Although all methods suffer as the number of objects increase, the proposed EZ-VSL consistently outperforms prior work.
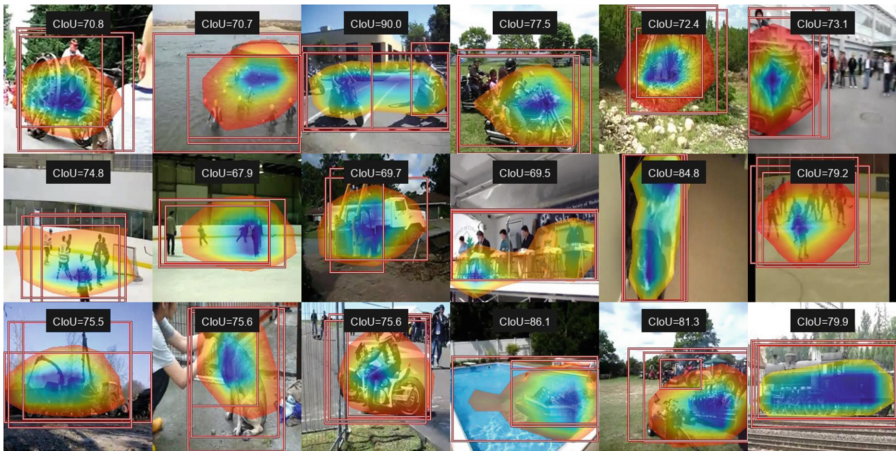
Fig. 6. Predicted localization maps on Flickr SoundNet test images.
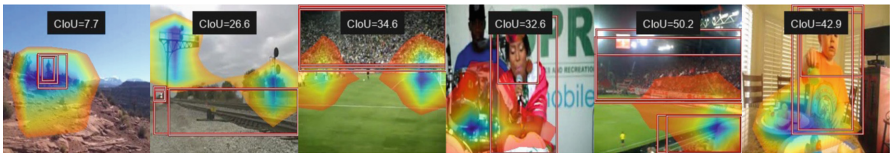
Fig. 7. Failure cases of EZ-VSL. Typical cases which EZ-VSL still struggles to accurately localize sound sources include small objects, or when sounds are not produced by objects, such as the sound of crowds.
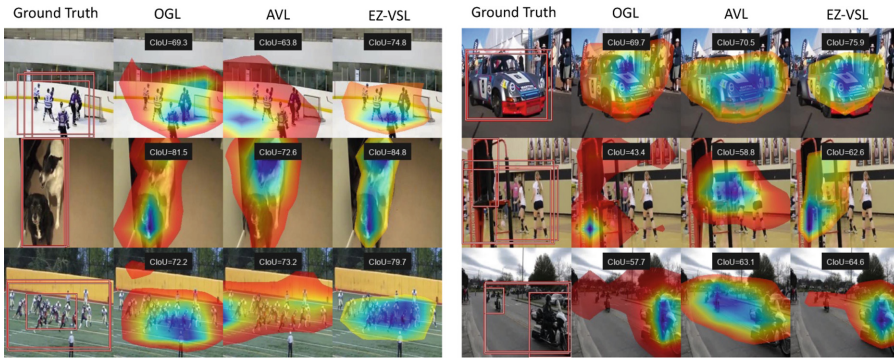
Fig. 8. Sound source localization by OGL, AVL, and our EZ-VSL. Object-guided maps tend to cover all objects in the scene; audio-video similarity maps often cover the sounding object and some non-object regions; the final EZ-VSL map tends to better focus on the sounding object.

5   Conclusion

In this work, we present the EZ-VSL, a simple yet effective approach for visual sounds source localization, with no need to explicitly compute the negative regions. Specifically, a simple cross-entropy loss is applied to learn the relative correspondence between the visual and audio instances. Furthermore, we propose a novel object-guided localization scheme to mix the audio-visual joint map and the object map from a lightweight pre-trained visual model for boosting the performance of orientating sound sources in an image. Compared to previous contrastive and non-contrastive baselines, our framework successfully achieves state-of-the-art performance on two popular benchmarks, Flickr SoundNet and VGG-Sound Source. Comprehensive ablation studies are conducted to show the effectiveness of each component in our simple method. We also demonstrate the significant advantage of our approach on the open set visual sounds source localization and cross dataset generalization.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: Deep lip reading: a comparison of models and an online application. In: Proceedings of Interspeech (2018)
2. Afouras, T., Owens, A., Chung, J.S., Zisserman, A.: Self-supervised learning of audio-visual objects from video. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12363, pp. 208–224. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58523-5_13
3. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 609–617 (2017)
4. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: learning sound representations from unlabeled video. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2016)

5. Chen, C., et al.: SoundSpaces: audio-visual navigation in 3D environments. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12351, pp. 17–36. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58539-6_2

6. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16867–16876 (2021)

7. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: VGGSound: a large-scale audio-visual dataset. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 721–725. IEEE (2020)

8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009)

9. Fisher III, J.W., Darrell, T., Freeman, W., Viola, P.: Learning joint statistical models for audio-visual fusion and segregation. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2000)

10. Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A.: Music gesture for visual sound separation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10478–10487 (2020)

11. Gan, C., Zhao, H., Chen, P., Cox, D., Torralba, A.: Self-supervised moving vehicle tracking with stereo sound. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7053–7062 (2019)

12. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 35–53 (2018)

13. Gao, R., Grauman, K.: 2.5D visual sound. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 324–333 (2019)

14. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3879–3888 (2019)

15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)

16. Hershey, J., Movellan, J.: Audio vision: using audio-visual synchrony to locate sounds. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (1999)

17. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9248–9257 (2019)

18. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 2127–2136 (2018)

19. Kidron, E., Schechner, Y.Y., Elad, M.: Pixels that sound. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2005)

20. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

21. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2018)

22. Morgado, P., Li, Y., Nvasconcelos, N.: Learning representations from audio-visual spatial alignment. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), pp. 4733–4744 (2020)
23. Morgado, P., Misra, I., Vasconcelos, N.: Robust audio-visual instance discrimination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12934–12945 (2021)
24. Morgado, P., Nvasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2018)
25. Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12475–12486, June 2021
26. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 631–648 (2018)
27. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 801–816. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_48
28. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12365, pp. 292–308. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58565-5_18
29. Rouditchenko, A., Zhao, H., Gan, C., McDermott, J.H., Torralba, A.: Self-supervised audio-visual co-segmentation. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2357–2361 (2019)
30. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4358–4366 (2018)
31. Senocak, A., Ryu, H., Kim, J., Kweon, I.S.: Learning sound localization better from semantically similar samples. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2022)
32. Son Chung, J., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6447–6456 (2017)
33. Tian, Y., Li, D., Xu, C.: Unified multisensory perception: weakly-supervised audio-visual video parsing. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12348, pp. 436–454. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58580-8_26
34. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4325–4334 (2017)
35. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1735–1744 (2019)
36. Zhao, H., et al.: The sound of pixels. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 570–586 (2018)