Tackling Overfitting and Silence in Unsupervised Audio-Visual Source Localization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Audio-visual source localization is a challenging task that aims to predict the location of visual sound sources in a video. Since collecting ground-truth annotations of sounding objects can be costly, a plethora of unsupervised localization methods have been proposed in recent years. Despite significant interest, current evaluation protocols have two major flaws. First, they allow for the use of a fully annotated dataset to perform early stopping, thus defeating the purpose of *unsupervised* source localization. Second, current evaluation metrics assume the presence of sound sources at all times. This is of course an unrealistic assumption, and thus better metrics are necessary to capture the model's performance on (negative) samples with no visible sound sources. To accomplish this, we extend the test set of popular benchmarks, Flickr SoundNet and VGG-Sound Sources, in order to include negative samples, and measure performance using metrics that balance precision and recall. Using the new protocol, we conducted an extensive evaluation of prior methods, and found that most prior works are not capable of identifying negatives and suffer from significant overfitting problems (rely heavily on early stopping for best results). We also propose a new approach for visual sound source localization that addresses both these problems. In particular, we found that, through extreme visual dropout and the use of momentum encoders, the proposed approach combats overfitting effectively, and establishes a new state-of-the-art performance on both Flickr SoundNet and VGG-Sound Source.

1 Introduction

From the meow of a cat to the roar of a lion to the revving of a motorcycle, humans have evolved to localize sources of sound in their environment to increase chances of survival. This ability relies in part on the uniqueness of sounds made by different entities – distinct sounds can be associated with their sources, which in turn can be visually localized in the environment. Due to the fast progress in audio and visual perception, several efforts have been made to develop multi-modal systems capable of mimicking our ability to visually localize sound sources.

One promising direction is to leverage the co-occurrence between sounds and the corresponding (visible) sources in video data. Since audio-visual co-occurrence arises naturally, algorithms can scale to very large datasets, without requiring costly human annotations. However, despite the significant efforts made to develop self-supervised localization methods [1, 2, 3, 4, 5], currently accepted evaluation protocols hide two critical limitations of current approaches: (1) current methods overfit easily even when scaled up to large datasets, and (2) current methods assume that visible sound sources are always present in the video and thus are unreliable when deployed on realistic data where this assumption does not hold.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

The first limitation remained hidden as prior works [2, 4, 6, 5] rely heavily on early stopping for optimal performance (*i.e.*, by continuously validating the model during training using a human-annotated test set). However, this practice not only violates the unsupervised assumption but is also at odds with the major promise of unsupervised learning – *i.e.*, enhanced capabilities when training on larger amounts of data for longer.

The second limitation remained hidden as current benchmark datasets [1, 6] and evaluation metrics only assess the model's ability to localize a sound source when one is present in the video. Importantly, it ignores a model's ability to correctly predict the *absence* of visual sound sources. This has led to a bias towards precision with a complete disregard for false positives. However, effective deployment of source localization in realistic settings requires both low type I and type II error rates.

The two limitations described above highlight the need for a more balanced and complete evaluation protocol for visual sound source localization. To achieve this, we extend popular benchmark test sets (Flickr SoundNet [1] and VGG-Sound Source [6]) to include 'negative' samples without any visible sound sources. We then conduct an extensive evaluation of existing methods [2, 3, 4, 6, 7], where in addition to overall precision, we also assess methods based on their ability to predict negative samples (through average precision and F1 scores). We observe that all prior work suffers from significant overfitting problems, and thus leans heavily upon early stopping to achieve the best performance. We also found that previous approaches fail to identify the frames where no visible sound sources are present.

To address these issues, we further propose a new algorithm - Momentum Audio-Visual Matching for Visual Sound Localization (`MoVSL`). First, to combat overfitting, we adopt slow-moving momentum audio and visual encoders and extreme visual dropout. Second, we localize visual sources by simultaneously performing sounding region selection and cross-modal instance discrimination. The first term highlights which regions within an image are most associated with particular audio, while the second term downplays regions that can be better described by the audio of other samples. By combining these two terms, the model is able to both accurately localize visible sources and identify when no visible sources are present. Using the newly developed evaluation protocol, we show that, unlike all prior work, `MoVSL` does not overfit and thus can be trained without relying on early stopping (which requires human annotations). `MoVSL` achieves state-of-the-art performance on multiple datasets and is more accurate at identifying samples without any visible sound sources.

2 Related Work

Audio-Visual Joint Learning. Audio-visual joint learning has been explored in many works [8, 9, 10, 11, 12, 1, 13, 14, 15, 16, 17, 18, 19, 20] to learn the temporal correlation between audio and visual information from videos. Given a sequence of videos, the main idea is to close the distance of audio and visual features from the same timestamp while pushing away those from different timestamps. Such audio-visual temporal synchronization is beneficial to several tasks, such as audio separation [17, 21, 22, 14, 15, 23, 24, 25], audio-visual spatialization [13, 26, 27, 18], visual sound source localization [1, 2, 3, 4, 6, 28]. In this work, we mainly focus on audio-visual source localization on videos without annotations involved during training, which requires us to learn meaningful audio-visual representations between a clip of audio and the frame centering around it.

Audio-Visual Source Localization. Audio-visual source localization aims at predicting the location of the sounding object in a video. Early machine learning and deep models [29, 30, 31] were used to capture the correspondences of audio and visual features. Recently, contrast learning based approaches [32, 2, 6, 7] have been proposed to learn the audio-visual alignment between features from two stream networks. For instance, LVS [6] leveraged a contrastive loss with hard negatives mining to learn the audio-visual co-occurrence map discriminatively. Contrastive learning with hard positives was introduced in HardPos [7] to learn audio-visual alignment with negative samples. More recently, EZ-VSL [5] proposed the multiple instance contrastive learning without positive and negative regions for training to match the most similar location in the corresponding image and distinguish all locations in different images.

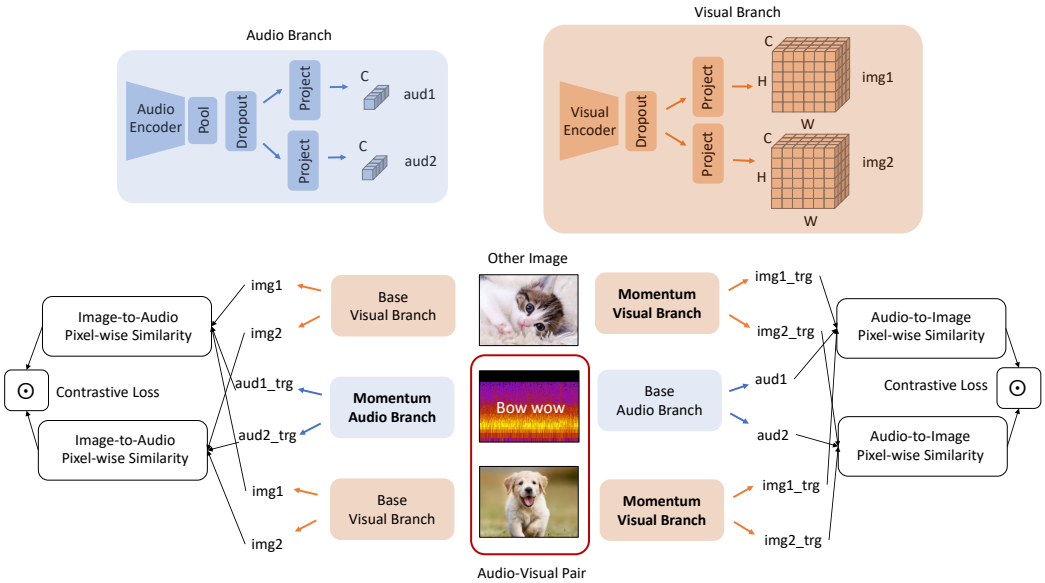

Figure 1: Illustration of the proposed MoVSL with synchronized momentum audio-visual matching. The audio-visual correspondence of pixel-wise similarity between momentum branches and base branches is aligned by the cross-modal multiple instance contrastive learning objective.

However, these baselines usually performed the model validation on the test set during training to select the best model as the final checkpoint, which caused the model to overfit. Furthermore, they only focus on localizing sound sources that are visible, ignoring non-visible or non-audible samples in real videos. To address these challenges, we propose a novel momentum audio-visual matching approach for learning the audio-visual correspondence between momentum and base features. We also explore both sounding and non-sounding frames in this work and propose new metrics for comprehensive evaluation.

3 Method

We propose a new Momentum-based approach for Visual Sound Localization, denoted `MoVSL`, by minimizing the image-to-audio/audio-to-image pixel-wise similarity between momentum audio/visual features and base visual/audio embeddings. In section 3.1, we define the problem and notations used in this paper and revisit multiple-instance contrastive learning proposed in EZ-VSL [5]. In section 3.2, we detail the proposed Momentum Audio-Visual Matching strategy. Finally, we detail the new evaluation protocol including the extended test sets for Flickr-SoundNet/VGG-SS and evaluation metrics in sections 3.3 and 3.3.

3.1 Preliminaries: Visual Source Localization by Localized Audio-Visual Matching [5]

Given an audio-visual paired dataset $\mathcal{D} = \{(v_i, a_i) : i = 1, \ldots, N\}$, visual source localization (VSL) aims at predicting the location of the sources present in sound a_i within the visual frame v_i. This is accomplished by matching the sound a_i to each region of v_i in a learned common space. To this end, an audio encoder $f_a(\cdot)$ is first used to extract global audio features, and a visual encoder $f_v(\cdot)$ to extract local visual features at each (x, y) location. Both features are then mapped into a common space using two projection heads $g_a(\cdot)$ and $g_v(\cdot)$ for the audio and visual features, respectively. This results in a set of visual embeddings spanning all locations in an image $V_i = \{\mathbf{v}_i^{xy} : \forall x, y\}, \forall i \in \mathcal{D}, x \in [1, W], y \in [1, H]$ and one global audio embedding $\mathbf{a}_i$.

All components of the model are then trained to align the audio embedding $\mathbf{a}_i$ with the visual embeddings $\mathbf{v}_i^{xy}$ at the locations (x, y) where the sound sources are located. If trained successfully, the model can be used for VSL by computing the audio-visual similarity at each location. However, for unsupervised VSL, the ground-truth locations of sound sources are not known during training.

To bypass the lack of annotations, prior work [5] proposed a multiple-instance contrastive learning objective, in which the model is trained to match audio representations $\mathbf{a}_i$ with at least one location in the positive set V_i, while pushing away from the visual features in all negative bags in the same mini-batch V_j $\forall j \neq i$ (at all locations). In particular, the similarity between an audio feature $\mathbf{a}_i$ and a bag of visual features $V_j = \{\mathbf{v}_j^{xy} : \forall x, y\}$ is computed by max-pooling audio-visual cosine similarities across spatial locations $\texttt{sim}(\mathbf{a}_i, V_j) = \max_{xy} s(\mathbf{a}_i, \mathbf{v}_j^{xy})$. The model is then trained to minimizing the loss function

$$\mathcal{L} = -\log \frac{\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_i, V_i)\right)}{\sum_j \exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_i, V_j)\right)} - \log \frac{\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_i, V_i)\right)}{\sum_j \exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{a}_j, V_i)\right)} \tag{1}$$

where τ is a temperature hyper-parameter.

3.2 Momentum Audio-visual Matching

We now introduce the proposed method, `MoVSL`, for visual sound localization. As we will show, most prior work, including the current state-of-the-art EZ-VSL [5], suffer from two critical limitations: (1) models can easily overfit to the self-supervised objective; (2) models are not capable of identifying negative samples (*i.e.*, samples with no visible sound sources).

The proposed method, illustrated in Figure 1, is designed to address these two limitations. To prevent overfitting, we apply two regularization techniques: dropout [33] and (slow-moving) momentum encoders [34, 35]. To enhance the identification of negatives, we propose to conduct VSL by simultaneously performing sounding region selection and instance discrimination.

Dropout We apply dropout [33] on the output of the f_v and f_a encoders. Interestingly, very heavy visual dropout is required to prevent overfitting (with a dropout probability as large as 0.9). Since multiple-instance contrastive learning (Eq. 1) matches the audio to only one of the $H \times W$ locations in the image, spurious alignment becomes more likely. Heavy visual dropout reduces the likelihood of spurious audio-visual alignments, as visual features become more spatially redundant [1]

Momentum encoders We use momentum audio-visual encoders [34, 35] $\hat{f}_a(\cdot), \hat{f}_v(\cdot)$ to compute (slow-moving) representations $\hat{\mathbf{a}}_i$ and $\hat{\mathbf{v}}_i^{xy}$. These momentum features are then used as more stable targets for audio-visual matching. Following [34, 35], the momentum encoders are updated using an exponential moving average of the corresponding online encoders with coefficient m.[2] Since updates to the audio-visual encoders are slowly incorporated into the momentum encoders, the target representations display a smoother behavior during the training process.

Enhancing negative detection with simultaneous localization and instance discrimination (SLID)

Inspired by prior work on weakly-supervised detection [36], we factorize VSL into two probability terms. The first is a localization term P^{loc} that seeks to distinguish within an image the regions that likely depict the sound source from the regions that likely do not. The second is a cross-modal instance discrimination term P^{xid} that seeks to highlight regions of an image that are likely associated with the audio, while suppressing regions that would be better explained by sound sources from other samples. The key insight is that to enhance negative detection, VSL should only select the regions that are *simultaneously* highlighted in both terms. While the first term is responsible for sound source localization, and the second prevents VSL to be overly confident when sound sources are not visible in the image.

To enable different interactions, the localization and instance discrimination terms are defined as audio-visual similarities computed on different spaces. In particular, let $\mathbf{a}_i^{\text{loc}} = g_a^{\text{loc}}(\mathbf{a}_i)$ and $\mathbf{a}_i^{\text{xid}} = g_a^{\text{xid}}(\mathbf{a}_i)$ be the audio representation of a_i in the two separate spaces, and $V_i^{\text{loc}} = \{g_v^{\text{loc}}(\mathbf{v}_i^{xy}) : \forall x, y\}$ and $V_i^{\text{xid}} = \{g_v^{\text{xid}}(\mathbf{v}_i^{xy}) : \forall x, y\}$ be the two bags of visual features for v_i. Here, the various functions $g(\cdot)$ denote simple linear projections. The localization term is computed as the softmax $\rho_{xy}(\cdot)$ of

[1] Spatial redundancy is encouraged as it prevents loss of information when features are dropped.

[2] EMA update is $\hat{\theta} \leftarrow \hat{\theta} + (1-m)\theta$, where θ and $\hat{\theta}$ denote the parameters of online and momentum functions.

localized audio-visual similarities over the spatial dimensions x, y

$$P^{\text{loc}}(a_i, v_j^{xy}) = \rho_{xy}\left(\frac{1}{\tau} s(\mathbf{a}_i^{\text{loc}}, g_v^{\text{loc}}(\mathbf{v}_j^{xy}))\right), \tag{2}$$

and the instance discrimination term as the softmax $\rho_i(\cdot)$ over instances i

$$P^{\text{xid}}(a_i, v_j^{xy}) = \rho_i\left(\frac{1}{\tau} s(\mathbf{a}_i^{\text{xid}}, g_v^{\text{xid}}(\mathbf{v}_j^{xy}))\right). \tag{3}$$

To select the regions in which both P^{loc} and P^{xid} are active, the two terms are combined into a single prediction map $P_{ij}^{xy} = P^{\text{loc}}(a_i, v_j^{xy}) \cdot P^{\text{xid}}(a_i, v_j^{xy})$, and the model is trained to optimize

$$\mathcal{L} = -\log \frac{\exp\left(\max_{xy} P_{ii}^{xy}\right)}{\sum_j \exp\left(\max_{xy} P_{ij}^{xy}\right)} - \log \frac{\exp\left(\max_{xy} P_{ii}^{xy}\right)}{\sum_j \exp\left(\max_{xy} P_{ij}^{xy}\right)}. \tag{4}$$

Finally, to combine SLID momentum encoders, audio and visual target representations are used when computing Eq. 2 and Eq. 3.

During inference, both localization and instance discrimination terms are used for source localization. We also observed that the (more stable) momentum features lead to improved visual source localization. Thus, given an audio-visual pair (v, a), the audio-visual localization map is computed as

$$\mathbf{S}_{xy}^{AVL} = s(\hat{\mathbf{v}}_{xy}^{\text{loc}}, \hat{\mathbf{a}}^{\text{loc}}) + s(\hat{\mathbf{v}}_{xy}^{\text{xid}}, \hat{\mathbf{a}}^{\text{xid}}). \tag{5}$$

Furthermore, this audio-visual localization map can be easily combined with object-guided localization proposed in [5] for improved performance.

3.3 Benchmarking Visual Source Localization

We introduce a new evaluation protocol for VSL.

Extended Flickr-SoundNet/VGG-SS Non-visible sounds or frames with silent objects can often be found in videos. To tackle these cases, we present a new evaluation protocol. We extended VGG-SS/Flickr-SoundNet by merging clips with no sounding objects to the original VGG-SS/Flickr-SoundNet test set. Specifically, we download 1k YouTube videos from the large-scale audio-visual dataset, VGG-Sound [37], and manually select 5-second clips with non-audible frames and/or non-visible sound sources. This resulted in 379 non-sounding (negative) samples.

Beyond the manually identified negative pairs, we further generate negative samples by pairing audio and videos that do not belong together. We control the difficulty of these negative pairs, by sampling 25% of pairs from audio and video that are from the same class (hard negative), and 75% from different classes. We merge all negatives with the VGG-SS [6] and Flickr-SoundNet [1] test set which only contain positive pairs. Table 1 shows the final statistics of the final test sets. We also split sounding samples into four groups (Small, Medium, Large, and Huge), and show the shows the distribution over sources of different sizes of the ground-truth bounding box.

Evaluation metrics Prior work [1] introduced consensus intersection over union (cIoU, denoted as u) for evaluating predicted localization maps, by comparing them to a group of human annotations. Given a set of predictions with cIoUs $\mathcal{U} = \{u_i\}_{i=1}^N$, each prediction is considered to be correct if its cIoU is above the cIoU threshold γ (usually set at γ). Thus, the Precision over all positive samples (*i.e.* with visible sounding objects) is computed as

$$\text{Precision} = \frac{|\mathcal{TP}|}{|\mathcal{TP}| + |\mathcal{FP}|} \quad \text{where} \quad \mathcal{TP} = \{i | u_i > \gamma\}, \qquad \mathcal{FP} = \{i | u_i \leq \gamma\} \tag{6}$$

Existing work [2, 3, 4, 6, 7, 5] usually reports a metric (misguidingly) called "CIoU", which is the Precision at $\gamma = 0.5$.

In this work, we also consider samples with no visible sounding sources. Thus, we consider a dataset $\mathcal{D} = \{(v_i, a_i, c_i) : i = 1, \ldots, N\}$, where c_i is 1 for samples with visible sources (positive samples) and 0 for samples with no visible sources (negative samples). Furthermore, in addition

Table 1: Statistics of benchmarks in unsupervised audio-visual source localization.

Dataset Size of Ground-Truth	Small $(0, 32^2]$	Medium $(32^2, 96^2]$	Large $(96^2, 144^2]$	Huge $(144^2, 224^2]$	Sounding $(0, 224^2]$	Non-Sounding 0	Total $[0, 224^2]$
Flickr-SoundNet [1]	0	9	83	158	250	0	250
Extended Flickr-SoundNet (ours)	0	9	83	158	250	**250**	**500**
VGG-Sound Source [6]	134	1796	1726	1502	5158	0	5158
Extended VGG-SS (ours)	134	1796	1726	1502	5158	**5158**	**10316**

to the localization map, the model is required to predict whether the current pair contains a visible source or not. This is accomplished by computing a confidence score d_i for each sample i which we define as the maximum value in the predicted audio-visual similarity map. Taking the confidence score into account, we formulate true positives as $\mathcal{TP}' = \{i|c_i = 1, d_i > \delta, u_i > \gamma\}$, false positives as $\mathcal{FP}' = \{i|c_i = 1, d_i > \delta, u_i \leq \gamma\} \cup \{i|c_i = 0, d_i > \delta\}$ and false negatives as $\mathcal{FN}' = \{i|c_i = 1, d_i \leq \delta\}$. Then, to comprehensively evaluate both sounding and non-sounding samples, we use Average Precision (AP) and maximum F1 (max-F1) score obtained by sweeping the confidence threshold δ (*i.e.* $\max_\delta F1(\delta)$). The detailed formula can be seen in Appendix. In experiments, we use $\gamma = 0.5$ for evaluation.

4 Experiments

In this section, we evaluate the proposed MoVSL on unsupervised audio-visual source localization. Following previous work [6, 7, 5], we use the VGG-Sound dataset [37] for training without annotations. Then our trained models are tested on four datasets: Flickr-SoundNet testset [1], VGG-Sound Source [6], the proposed Extended Flickr-SoundNet and Extended VGG-SS. We report the quantitive results using Precision-Visible on datasets with only sounding samples. AP and max-F1 are reported for benchmarks with both sounding and non-sounding samples.

4.1 Experimental setup

Datasets. VGG-Sound [37] contains 200k videos of 309 sound categories from YouTube. For unsupervised learning, we extract audio-visual pairs with a single frame and its corresponding audio of 20 seconds long. Following commonly-used settings [6, 7, 5], we use subset of 144k pairs for training. We test our model on two types of datasets: 1) with only sounding samples: Flickr-SoundNet testset [1] and VGG-Sound Source [6]; 2) with both sounding and non-sounding samples: the proposed Extended Flickr-SoundNet and Extended VGG-SS. The Flickr-SoundNet includes only 250 audio-visual pairs of sounding objects, while VGG-Sound Source contains 5158 pairs of sounding objects from 220 categories. Extended Flickr-SoundNet totally includes 500 samples, 250 sounding and 250 non-sounding. Extended VGG-SS spans 5158 pairs of sounding objects and 5158 pairs of non-sounding objects, 10,316 in total.

Implementation. For audio and visual encoders, we use ResNet-18 [38] as the backbone, and initialize the visual encoder with ImageNet [39] pre-trained weights, as previous work [6, 7, 5] did. The dimension of audio-visual features in the projected share space is 512. For visual input, we first resize frames of 224×224 to 246 along the shortest edge, and then apply random cropping with random horizontal flipping on them at training time, We do not use any data augmentation and use resized frames of 224×224 for inference. For audio encoder, we extract the log spectrograms with 257 frequency bands over 300 timesteps from 3s of audio as input. Following the prior work [5], we train the model with a batch size of 128 on 2 GPUs for 20 epochs. The Adam [40] optimizer with β_1=0.9, β_2=0.999 is applied with an initial learning rate of 1e-4 and weight decay of 1e-4. The probability of dropout for visual branch is 0.9. The momentum coefficients m_a=0.999, m_v=0.999.

4.2 Comparison of overfitting to prior work

To demonstrate the effectiveness of our model in alleviating the overfitting issue, we compare both best and latest checkpoint with prior work [1, 2, 3, 4, 6, 7, 5] using the common metric Precision-Visible, that is, "CIoU". Table 2 shows the comparison results on Flickr SoundNet and VGG-SS testsets.

Table 2: Comparison results of Precision-Visible ("CIoU") for BEST and LATEST checkpoints on Flickr SoundNet and VGG-SS testsets where models are trained on VGG-Sound 144k data. 'OGL' denote the object-guided localization proposed in EZ-VSL [5]. ⋆ denotes the value reported in the LVS [6] and HardPos[7] papers.

Method	Flickr-SoundNet		VGG-SS	
	BEST	LATEST	BEST	LATEST
Attention10k [1]	42.26	34.16	18.50⋆/18.52	14.04
CoarsetoFine [4]	–	47.20	–	21.93
DMC [2]	55.60	52.80	23.90	22.63
AVObject [3]	–	–	29.70⋆	–
LVS [6]	71.90⋆/71.60	19.60	34.40⋆/33.36	10.43
HardPos [7]	76.80⋆	–	34.60⋆	–
EZ-VSL [5]	79.60	66.40	34.28	31.58
EZ-VSL + OGL [5]	83.94⋆/83.94	72.80	38.85⋆/38.85	37.86
MoVSL (ours)	**83.20**	**83.60**	**37.22**	**37.79**
MoVSL (ours) + OGL [5]	**86.40**	**86.00**	**39.67**	**39.80**

Table 3: Comparison results of the proposed metrics (AP, max-F1, Precision) on Extended Flickr-SoundNet and Extended VGG-SS benchmark where models are trained on VGG-Sound 144k data.

Method	Extended Flickr-SoundNet			Extended VGG-SS		
	AP	max-F1	Precision	AP	max-F1	Precision
Center Prior	–	–	67.60	–	–	34.16
CoarsetoFine [4]	0.00	38.20	47.20	0.00	19.80	21.93
LVS [6]	9.80	17.90	19.60	5.15	9.90	10.43
Attention10k [1]	15.98	24.00	34.16	6.70	13.10	14.04
DMC [2]	25.56	41.80	52.80	11.53	20.30	22.63
OGL [5]	40.20	55.70	77.20	18.73	30.90	36.58
EZ-VSL [5]	46.30	54.60	66.40	24.55	30.90	31.58
EZ-VSL + OGL [5]	48.75	56.80	72.80	27.71	34.60	37.86
MoVSL (ours)	**51.63**	**59.10**	**83.60**	**32.95**	**40.00**	**37.79**
MoVSL (ours) + OGL [5]	**52.15**	**60.10**	**86.00**	**34.46**	**41.50**	**39.80**

We can observe that our MoVSL superiorly outperforms previous work by using the best checkpoint for comparison, demonstrating the state-of-the-art results of our approach in this setting. Without the OGL (object-guided localization) proposed in EZ-VSL [5], we achieve the performance gains of 3.60% and 2.94% on the Flickr and VGG-SS tests sets. When the OGL is involved at inference time, our MoVSL further improves the SOTA performance by 2.46% and 0.82%. Significant gains can be observed in the setting of using the latest checkpoint for comparison. On the Flickr test set, our framework outperforms the state-of-the-art work by large margins, 17.20% and 13.20% for w/o OGL and w OGL. The performance gains of 6.21% and 1.94% on the challenging VGG-SS dataset are achieved, which shows the importance of our momentum audio-visual matching strategy in preventing the model from overfitting.

4.3 Benchmarking on Extended Flickr-SoundNet and VGG-SS

With the proposed metrics (AP, max-F1, Precision-Visible) in Sec 3.3, we benchmark state-of-the-art work [6, 5] and our model on two datasets with both sounding and non-sounding samples: Extended Flickr-SoundNet, and Extended VGG-SS. Note that the Center Prior is to generate a random Gaussian map in the middle of frames for predictions. The experimental results are reported in Table 3. As can be seen, the proposed MoVSL establishes new state-of-the-art performance in terms of all metrics. On the Extended Flickr-SoundNet with 250 additional non-sounding samples, we achieve 51.63% AP and 59.10% max-F1, but a high precision of 83.60% on visible objects. This shows the importance of taking false negatives into account for non-sounding samples. Due to the large amount (5158) of negatives in the Extended VGG-SS, AP and max-F1 are decreased to 32.95% and 40.00%, indicating the advantage of the proposed AP and max-F1 in evaluating non-sounding objects over the original

Table 4: Ablation study of MoVSL. We ablate the use of Momentum Encoders (MomEnc), SLID, and the inference procedure (Inf).

MomEnc	SLID	Infer	AP	max-F1	Precision
		$s(\mathbf{v}_{xy}, \mathbf{a})$	24.55	30.90	31.58
✓		$s(\hat{\mathbf{v}}_{xy}, \hat{a})$	28.19	35.60	32.75
✓	✓	$s(\hat{\mathbf{v}}^{\text{loc}}_{xy}, \hat{\mathbf{a}}^{\text{loc}})$	22.01	30.80	23.65
✓	✓	$s(\hat{\mathbf{v}}^{\text{xid}}_{xy}, \hat{\mathbf{a}}^{\text{xid}})$	17.29	29.60	34.72
✓	✓	Formula. 5	**32.95**	**40.00**	**37.79**

Table 5: Regularization techniques of weight decay (wd), visual dropout (vdrop), momentum (mom).

wd	vdrop	mom	AP	max-F1	Precision
–	–	–	32.32	39.30	39.74
1e-4	–	–	27.86	33.70	38.02
1e-4	0.9	–	28.45	36.30	37.86
1e-4	–	0.999	27.13	33.40	38.60
1e-4	0.9	0.999	**34.46**	**41.50**	**39.80**

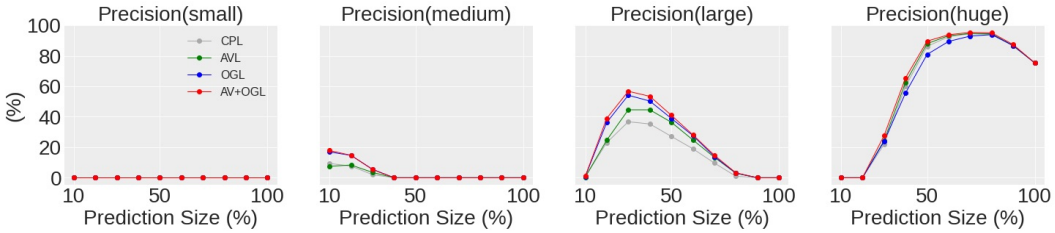

Figure 2: Effect of different relative sorting thresholds on Precision for visible objects of various sizes (Small, Medium, Large, Huge). CPL, AVL, OGL, and AV+OGL denote the localization using Gaussian center prior map, audio-visual response map, object-guided map, and the linear combination of audio-visual response and object-guided map.

metric, Precision-Visible. In the meanwhile, quantitative results of AP and max-F1 are positively correlated to those of Precision-Visible, which further implies the rationality of the proposed metrics in scoring both sounding and non-sounding samples.

4.4 Experimental analysis

In this section, we conduct an extensive benchmarking study of prior work. We train several methods [1, 4, 2, 3, 6, 7, 5] to evaluate them using the newly proposed evaluation protocol. When implementing prior work, we used authors implementation with default parameters. When not available, we carefully followed the paper description. We report both our reproduced results, and the results reported in the original paper, whenever possible.

To demonstrate the benefits of MoVSL, we compare our method to prior work using the new evaluation protocol. All models were trained on the VGG-Sound 144k subset and the latest checkpoints were tested on the challenging Extended VGG-SS with both sounding and non-sounding objects.

SLID localization strategy and regularization. In order to demonstrate the effectiveness of MoVSL, we ablated the use of momentum encoders and simultaneous localization and instance discrimination (SLID) strategies during training. Table 4 reports the experimental results. Adding momentum encoders to the baseline improves AP and max-F1 by large margins, 3.64% and 4.70%. This shows the effectiveness of the proposed momentum audio-visual matching approach in learning the cross-modal alignment between momentum and base representations. When SLID is used, we further increase all metrics by significant margins, 4.76% AP, 4.40% max-F1, and 5.04% Precision.

Furthermore, we ablated audio-visual matching strategies at inference time for the proposed MoVSL trained with the SLID objective. We evaluate the localization ability of each SLID branch separately, *i.e.* using the similarities between the global audio feature $\hat{\mathbf{a}}^{\text{loc}}$ ($\hat{\mathbf{a}}^{\text{xid}}$) and the bag of localized visual features $\hat{\mathbf{v}}^{\text{loc}}_{xy}$ ($\hat{\mathbf{v}}^{\text{xid}}_{xy}$). We compare these with the localization performance using both SLID branches as in Equation 5. Results are shown in Table 4. Both alternative approaches achieve worse results than our MoVSL which uses both SLID branches for localization. This further indicates the effectiveness of the proposed approach in capturing the cross-modal pixel-wise correspondence between momentum and base features. We also ablate the regularization techniques (weight decay, visual dropout, momentum) in Table 5. It can be seen that dropout in the visual feature is very important to prevent overfitting. The optimal momentum is 0.999 for both visual and audio models.

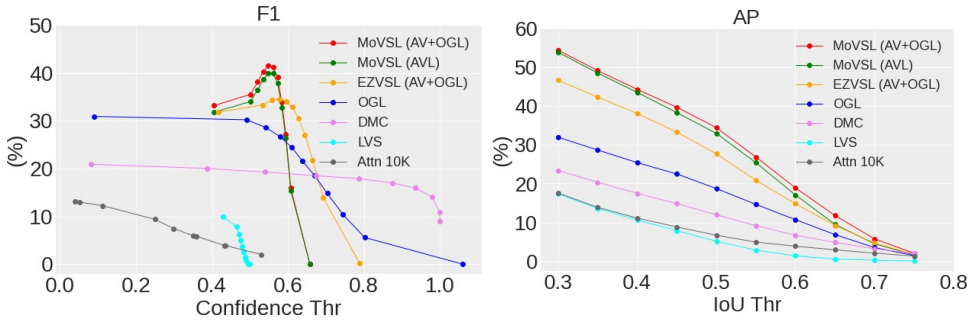

Figure 3: Effect of confidence thresholds on the F1 score and IoU thresholds on AP.

Effect of Precision, F1 score, and AP. To explore the effect of various prediction sizes on the Precision of visible objects within each subset (Small, Medium, Large, Huge), we vary the size from 10% to 100% and compare the performance using the CPL, AVL, OGL, AV+OGL for inference. The results are shown in Figure 2. All methods achieve the peak of Precision for medium, large, and huge objects at 10%, 30%, 70%. Due to the difficulty of identifying small objects in the whole frame, Precision for different thresholds is always 0. This highlights one important limitation of current unsupervised visual sound source localization methods. We also evaluated the F1 score and AP with different confidence and IoU thresholds. Figure 3 shows the comparison results of the proposed MoVSL and previous work. When the confidence is 0.5, our method achieves the highest F1 score against existing methods. Our method also outperforms all prior works in terms of AP using all IoU thresholds.

4.5 Limitations

Although we achieve superior precision on large and huge objects in the image, the performance of small and medium objects is not satisfactory, as shown in Figure 2. When it comes to small objects, the proposed method performs badly as the prior work did. The possible reason is that all methods use the lightweight ResNet-18 as the visual backbone to extract features, which limits the quality of visual representations for audio-visual matching. In the meanwhile, we notice that the prediction map is generated by the bilinear interpolation of an audio-visual response map with a low resolution. Increasing the resolution of such a map could be beneficial to small objects in the raw image with high resolution. Therefore, applying a strong backbone would be the future direction for improving the precision of small objects.

5 Conclusion

In this work, we explore two main issues in common early stopping settings and the visible sound source assumption for unsupervised audio-visual source localization. To capture the model's performance on samples with no visible sound sources, we extend the test sets of popular benchmarks, Flickr SoundNet and VGG-Sound Source, to cover negative samples, and use average precision and F1 scores as metrics to balance precision and recall. Extensive evaluations of prior work are performed under the new proposed evaluation protocol. We observe that most previous work suffers from significant overfitting problems, and thus relies heavily on early stopping for best results. To eliminate this, we propose a novel Momentum audio-visual matching approach for Visual Sound Localization by applying momentum features as targets of base embeddings for each modality, which we denote MoVSL. Thorough experiments demonstrate the effectiveness of the proposed approach in combating overfitting, and the state-of-the-art performance on two popular benchmarks, Flickr SoundNet and VGG-Sound Source.

References

[1] 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 (CVPR)*, pages 4358–4366, 2018.

[2] 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 (CVPR)*, pages 9248–9257, 2019.

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

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

[5] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022.

[6] 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, 2021.

[7] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

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

[9] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–816, 2016.

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

[11] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4325–4334, 2017.

[12] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[13] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360°video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[14] 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.

[15] 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 (ICCV)*, pages 1735–1744, 2019.

[16] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 436–454, 2020.

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

[18] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 4733–4744, 2020.

[19] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12934–12945, 2021.

[20] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12475–12486, June 2021.

[21] 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.

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

[23] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10457–10467, 2020.

[24] 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, 2021.

[25] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15495–15505, 2021.

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

[27] Changan Chen, Unnat Jain, Carl Schissler, S. V. A. Garí, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 17–36, 2020.

[28] Ziyang Chen, David F. Fouhey, and Andrew Owens. Sound localization by self-supervised time delay estimation. *arXiv preprint arXiv:2204.12489*, 2022.

[29] John Hershey and Javier Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 1999.

[30] John W Fisher III, Trevor Darrell, William Freeman, and Paul Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2000.

[31] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[32] 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.

[33] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[34] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[35] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[36] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2854, 2016.

[37] 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.

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

[39] Jia Deng, Wei Dong, Richard Socher, Li-Jia. Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Section 4.5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section **??**.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] See Supplementary material.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Training and test sets are too large to run each experiment multiple times. However, given their sizes, error bars are expected to be small.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.1.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]