Localizing Visual Sounds the Hard Way

Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani
Andrea Vedaldi, Andrew Zisserman
{hchen, weidi, afourast, arsha, vedaldi, az}@robots.ox.ac.uk
VGG, Department of Engineering Science, University of Oxford, UK
http://www.robots.ox.ac.uk/~vgg/research/lvs/

Abstract

The objective of this work is to localize sound sources that are visible in a video without using manual annotations. Our key technical contribution is to show that, by training the network to explicitly discriminate challenging image fragments, even for images that do contain the object emitting the sound, we can significantly boost the localization performance. We introduce a mechanism to mine hard samples and add them to a contrastive learning formulation automatically. We show that our algorithm achieves state-of-the-art performance on the popular Flickr SoundNet dataset. Furthermore, we introduce the VGG-Sound Source (VGG-SS) benchmark, a new set of annotations for the recently-introduced VGG-Sound dataset, where the sound sources visible in each video clip are explicitly marked with bounding box annotations. This dataset is 20 times larger than analogous existing ones, contains 5K videos spanning over 200 categories, and, differently from Flickr SoundNet, is video-based. On VGG-SS, we also show that our algorithm achieves state-of-the-art performance against several baselines.

1. Introduction

In this paper, we consider the problem of localizing 'visual sounds', *i.e*. visual objects that emit characteristics sounds in videos. Inspired by prior works [2, 8, 14], we formulate this as finding the correlation between the visual and audio streams in videos. These papers have shown that not only can this correlation be learned successfully, but that, once this is done, the resulting convolutional neural networks can be 'dissected' to localize the sound source spatially, thus imputing it to a specific object. However, other than in the design of the architecture itself, there is little in this prior work meant to improve the localization capabilities of the resulting models. In particular, while several models [1, 2, 14] do incorporate a form of spatial attention which should also help to localize the sounding object as a byproduct, these may still fail to provide a good *coverage* of the object, often detecting too little or too much of it.

In order to address this issue, we propose a new training scheme that explicitly seeks to spatially localize sounds in video frames. Similar to object detection [17], in most cases only a small region in the image contains an object of interest, in our case a 'sounding' object, with the majority of the image often being 'background' which is not linked to the sound. Learning accurate object detectors involves explicitly seeking for these background regions, prioritizing those that could be easily confused for the object of interest, also called *hard negatives* [6, 7, 11, 13, 15, 17].

In order to incorporate hard evidence in our unsupervised (or self-supervised) setting, we propose an automatic background mining technique through differentiable thresholding, *i.e*. regions with low correlation to the given sound are incorporated into a negatives set for contrastive learning. We show that this simple change significantly boosts sound localization performance on standard benchmarks, such as Flickr SoundNet [14].

To further assess sound localization algorithms, we also introduce a new benchmark (VGG-SS), based on the recently-introduced VGG-Sound dataset [4],

2. Method

Our goal is to localize objects that make characteristic sounds in videos, without using any manual annotation. Similar to prior work [2], we use a two-stream network to extract visual and audio representations from unlabelled video. For localization, we compute the cosine similarity between the audio representation and the visual representations extracted convolutionally at different spatial locations in the images. In this manner, we obtain a positive signal that pulls together sounds and relevant spatial locations. For learning, we also need an opposite negative signal. A weak one is obtained by correlating the sound to locations in other, likely irrelevant videos. Compared to prior work [1, 2], our key contribution is to *also* explicitly seek for hard negative locations that contain background or non-sounding objects in the *same* images that contain the sounding ones, leading to more selective and thus precise

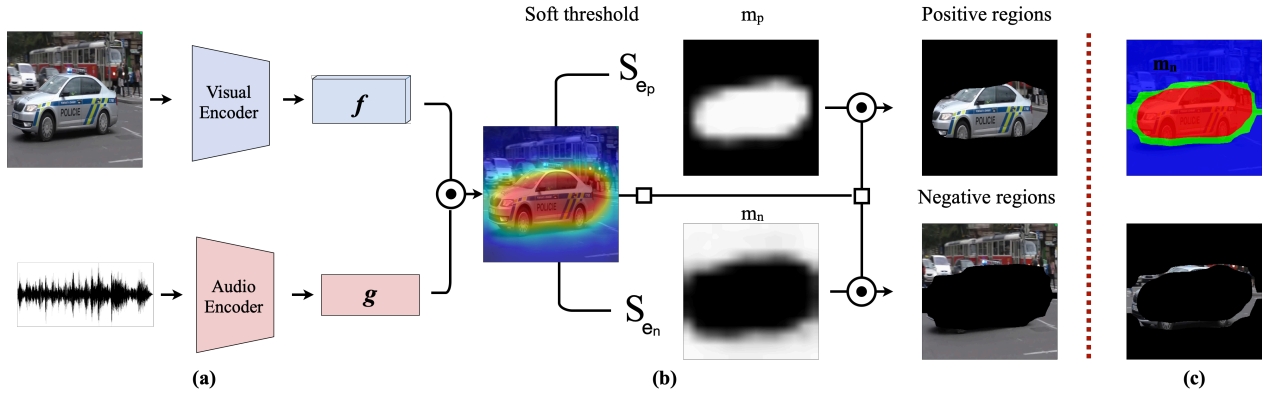


Figure 1: **Architecture Overview**. We use an audio-visual pair as input to a dual-stream network shown in **(a)**, $f(\cdot;\theta_1)$ and $g(\cdot;\theta_2)$, denoting the visual and audio feature extractor respectively. Cosine similarity between the audio vector and visual feature map is then computed, giving us a heatmap of size 14×14. **(b)** demonstrates the soft threshold being applied twice with different parameters, generating positive, negative regions. The final Tri-map are highlighed in **(c)**.

localization. An overview of our architecture can be found in Figure 1.

2.1. Audio-Visual Feature Representation

Given a short video clip with N visual frames and audio, and considering the center frame as visual input, *i.e.* $X = \{I, a\}, I \in \mathbb{R}^{3\times H_v \times W_v}, a \in \mathbb{R}^{1\times H_a \times W_a}$. Here, I refers to the visual frame, and a to the spectrogram of the raw audio waveform. In this manner, representations for both modalities can be computed by means of CNNs, which we denote respectively $f(\cdot;\theta_1)$ and $g(\cdot;\theta_2)$. For each video X_i, we obtain visual and audio representations:

$$V_i = f(I_i;\theta_1), \quad V_i \in \mathbb{R}^{c\times h\times w}, \tag{1}$$

$$A_i = g(a_i;\theta_2), \quad A_i \in \mathbb{R}^{c}. \tag{2}$$

2.2. Audio-Visual Correspondence

Given the video and audio representations of eqs. (1) and (2), we put in correspondence the audio of clip i with the image of clip j by computing the cosine similarity of the representations, using the audio as a probe vector:

$$[S_{i\to j}]_{uv} = \frac{\langle A_i, [V_j]_{:uv}\rangle}{\|A_i\|\,\|[V_i]_{:uv}\|}, \quad uv \in [h]\times[w].$$

2.3. Self-supervised Audio-Visual Localization

In this section, we describe a simple approach to continuously bootstrap the model to achieve better localization results. At a high level, the proposed idea inherits the spirit of self-training, where predictions are treated as pseudo-ground-truth for re-training.

Specifically, given a dataset $\mathcal{D} = \{X_1, X_2, \ldots, X_k\}$ where only audio-visual pairs are available (but not the masks m_i). To get the pseudo-ground-truth mask $\hat{m}_i$, we could simply threshold the map $S_{i\to i}$:

$$\hat{m}_i = \begin{cases} 1, & \text{if } S_{i\to i} \geq \epsilon \\ 0, & \text{otherwise} \end{cases}$$

Clearly, however, this thresholding, which uses the Heaviside function, is not differentiable. Next, we address this issue by relaxing the thresholding operator.

Smoothing the Heaviside function. Here, we adopt a smoothed thresholding operator in order to maintain the end-to-end differentiability of the architecture:

$$\hat{m}_i = \text{sigmoid}((S_{i\to i} - \epsilon)/\tau)$$

where ϵ refers to the thresholding parameter, and τ denotes the temperature controlling the sharpness.

Handling uncertain regions. The pseudo-ground-truth obtained from the model prediction may potentially be noisy, we therefore propose to set up an "ignore" zone between the positive and negative regions, allowing the model to self-tune. In the image segmentation literature, this is often called a Tri-map and is also used for matting [5, 16]. Conveniently, this can be implemented by applying two different ϵ's, one controlling the threshold for the positive part and the other for the negative part of the Tri-map.

Training objective. while computing the positives and negatives automatically, our final formulation are:

$$\hat{m}_{ip} = \text{sigmoid}((S_{i\to i} - \epsilon_p)/\tau)$$
$$\hat{m}_{in} = \text{sigmoid}((S_{i\to i} - \epsilon_n)/\tau)$$
$$P_i = \frac{1}{|\hat{m}_{ip}|}\langle \hat{m}_{ip},\ S_{i\to i}\rangle$$
$$N_i = \frac{1}{|\mathbf{1} - \hat{m}_{in}|}\langle \mathbf{1} - \hat{m}_{in},\ S_{i\to i}\rangle + \frac{1}{hw}\sum_{j\neq i}\langle \mathbf{1},\ S_{i\to j}\rangle$$
$$\mathcal{L} = -\frac{1}{k}\sum_{i=1}^{k}\left[\log\frac{\exp(P_i)}{\exp(P_i)+\exp(N_i)}\right]$$

where ϵ_p and ϵ_n are two thresholding parameters (validated in experiment section), with $\epsilon_p > \epsilon_n$. For example if we set $\epsilon_p = 0.6$ and $\epsilon_n = 0.4$, regions with correspondence scores above 0.6 are considered positive and bellow 0.4 negative, while the areas falling within the $[0.4, 0.6]$ range are treated as "uncertain" regions and ignored during training.

3. Experiments

In the following sections, we describe the datasets, evaluation protocol and experimental details used to thoroughly assess our method.

3.1. Training Data

For training our models, we consider two large-scale audio-visual datasets, the widely used Flickr SoundNet dataset and the recent VGG-Sound dataset, as detailed next. Only the center frames of the *raw* videos are used for training. Note, other frames *e.g.* (3/4 of the video) are tried for training, no considerable performance change is observed.

Flickr SoundNet: This dataset was initially proposed in [3] and contains over 2 million unconstrained videos from Flickr. For a fair comparison with recent work [10, 12, 14], we follow the same data splits, conducting self-supervised training with subsets of 10k or 144k image and audio pairs.

VGG-Sound: VGG-Sound was recently released with over 200k clips for 309 different sound categories. The dataset is conveniently audio-visual, in the sense that the object that emits sound is often visible in the corresponding video clip, which naturally suits the task considered in this paper. Again, to draw fair comparisons, we conduct experiments with training sets consisting of image and audio pairs of varying sizes, *i.e.* 10k, 144k and the full set.

3.2. Evaluation protocol

In order to quantitatively evaluate the proposed approach, we adopt the evaluation metrics used in [12, 14]: Consensus Intersection over Union (cIoU) and Area Under Curve (AUC) are reported for each model on two test sets, as detailed next.

Flickr SoundNet Testset: Following [10, 12, 14], we report performance on the 250 annotated image-audio pairs of the Flickr SoundNet benchmark.

VGG-Sound Source (VGG-SS): We also re-implement and train several baselines on VGG-Sound and evaluate them on our proposed VGG-SS benchmark, a new testing audio-visual localization benchmark with more than 5k videos spanning more than 200 classes.

3.3. Implementation details

Audio inputs are 257×300 magnitude spectrograms. The dimensions for the audio output from the audio encoder CNN is a 512D vector, which is max-pooled from a feature map of $17 \times 13 \times 512$, where 17 and 13 refer to the frequency and time dimension respectively. For the visual input, we resize the image to a $224 \times 224 \times 3$ tensor without cropping. For both the visual and audio stream, we use a lightweight ResNet18 [9] as a backbone. We use $\epsilon_p = 0.65$ and $\epsilon_n = 0.4$, $\tau = 0.03$, that are picked by ablation study.

4. Results

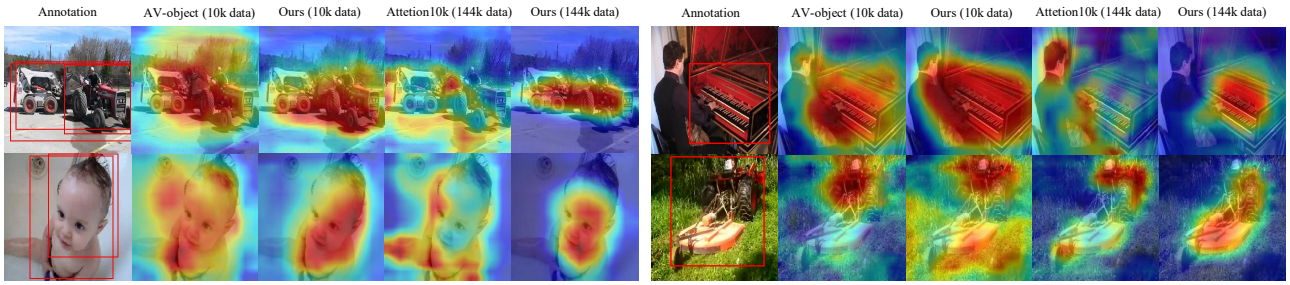
In the following sections, we first compare our results with recent work on both Flickr SoundNet and VGG-SS dataset in detail. Then we conduct an ablation analysis showing the importance of the *hard negatives* and the Tri-map in self-supervised audio-visual localization.

4.1. Comparison on the Flickr SoundNet Test Set

In this section, we compare to recent approaches by training on the same amount of data (using various different datasets). As shown in Table 1, we first fix the training set to be Flickr SoundNet with 10k training samples and compare our method with [2, 8, 12]. Our approach clearly outperforms the best previous methods by a substantial gap (0.546 vs. 0.582). Second, we also train on VGG-Sound using 10k random samples, which shows the benefit of using VGG-Sound for training. Third, we switch to a larger training set consisting of 144k samples, which gives us a further 5% improvement compared to the previous state-of-the-art method [10]. In order to tease apart the effect of various factors in our proposed approach, *i.e.* introducing *hard negative* and using a Tri-map vs different training sets, *i.e.* Flickr144k vs. VGG-Sound144k, we conduct an ablation study, as described next.

4.2. Comparison on VGG-Sound Source

In this section, we evaluate the models on the newly proposed VGG-SS benchmark. As shown in Table 2, the CIoU is reduced significantly for all models compared to the results in Table 1, showing that VGG-SS is a more diverse and challenging benchmark than Flickr SoundNet. How-



(a) Visualisation on Flickr SoundNet testset

(b) Visualisation on VGG-SS testset

Figure 2: **Qualitative results** for models trained on various methods and data amount. The first column shows annotation overlaid on images, the following two column shows predictions trained on 10k data and the last tow column show predictions trained on 144k data. Our method has no false positives in the predictions as the hard negatives are penalised in the training.

Method	Training set	CIoU	AUC
Attention10k [14]	Flickr10k	0.436	0.449
CoarsetoFine [12]	Flickr10k	0.522	0.496
AVObject [1]	Flickr10k	0.546	0.504
Ours	Flickr10k	**0.582**	**0.525**
Ours	VGG-Sound10k	**0.618**	**0.536**
Attention10k [14]	Flickr144k	0.660	0.558
DMC [10]	Flickr144k	0.671	0.568
Ours	Flickr144k	**0.699**	**0.573**
Ours	VGG-Sound144k	**0.719**	**0.582**
Ours	VGG-Sound Full	**0.735**	**0.590**

Table 1: Quantitative results on Flickr SoundNet testset. We outperform all recent works using different training sets and number of training data.

Method	CIoU	AUC
Attention10k [14]	0.185	0.302
AVobject [1]	0.297	0.357
Ours	**0.344**	**0.382**

Table 2: Quantitative results on the VGG-SS testset. All models are trained on VGG-Sound 144k and tested on VGG-SS.

ever, our proposed method still outperforms all other baseline methods by a large margin of around 5%.

4.3. Qualitative results

We visualize the prediction results in Figure 2, and note that the proposed method presents much cleaner heatmap outputs. This once again indicates the benefits of considering hard negatives during training.

5. Conclusion

We revisit the problem of unsupervised visual sound source localization and introduce a new large-scale benchmark called VGG-Sound Source. We also suggest a simple, general and effective technique that significantly boosts the performance of existing sound source locators, by explicitly mining for hard negative image locations in the same image that contains the sounding objecs. A careful implementation of this idea using Tri-maps and differentiable thresholding allows us to significantly outperform the state of the art.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proc. ECCV*, 2020. 1, 4

[2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proc. ECCV*, 2018. 1, 3

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*, 2016. 3

[4] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGG-Sound: A large-scale audio-visual dataset. In *Proc. ICASSP*, 2020. 1

[5] Yung-Yu Chuang, Aseem Agarwala, Brian Curless, David H. Salesin, and Richard Szeliski. Video matting of complex scenes. *ACM Trans. Graph*, 2002. 2

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005. 1

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014. 1

[8] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proc. ECCV*, 2018. 1, 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 3

[10] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clus-

tering for unsupervised audiovisual learning. In *Proc. CVPR*, June 2019. 3, 4

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. ICCV*, 2017. 1

[12] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proc. ECCV*, 2020. 3, 4

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2016. 1

[14] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proc. CVPR*, 2018. 1, 3, 4

[15] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proc. CVPR*, 2016. 1

[16] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proc. CVPR*, 2018. 2

[17] Paul Viola and Michael Jones. Robust real-time object detection. In *Proc. SCTV Workshop*, 2001. 1