Benchmarking Weakly-Supervised Audio-Visual Sound Localization

Shentong Mo[1] and Pedro Morgado[2]

[1] Carnegie Mellon University
[2] University of Wisconsin-Madison

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 weakly-supervised localization methods that can learn from datasets with no bounding-box annotations have been proposed in recent years, by leveraging the natural co-occurrence of audio and visual signals. Despite significant interest, popular evaluation protocols have one major flaw. That is, 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, to include negative samples, and measure performance using metrics that balance localization accuracy 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.

1 Introduction

Humans and most other animals have evolved to localize sources of sound in their environment. This remarkable ability relies in part on the uniqueness of different sound sources, which allows us to recognize the sounds we hear and visually localize them in our environment. Given the significant advances in audio and visual perception [12,2,10,9,8], there is broad interest in developing 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 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 encouraging recent progress [13,5,1,11,7], currently accepted evaluation protocols hide two critical limitations: (1) current methods overfit easily even when scaled 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.

The first limitation remained hidden as prior works [5,11,3,7] rely heavily on early stopping for optimal performance (*i.e.*, by continuously validating the model during training using a human-annotated set). The second limitation remained hidden as most prominent benchmark datasets [13,3] and evaluation metrics only assess the ability to localize sound sources when one is present in the video. Importantly, it ignores the ability to correctly predict the *absence* of visual sound sources. This has led to a bias towards localization accuracy with disregard for false positive detection.

The limitations 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 [13] and VGG-Sound Sources [3]) to include 'negative' samples without any visible sound sources. We conduct an extensive evaluation of existing methods [5,1,11,3,14,6], where in addition to overall localization accuracy, we also assess methods based on their ability to predict negative samples. We found that previous approaches do indeed overfit easily and struggle to strike a good balance between false positive and false negative rates. Evaluation code and extended test sets will be publicly available. These challenges are also addressed in a new framework presented in `https://github.com/stoneMo/SLAVC`.

2 Benchmarking Visual Source Localization

We introduce an evaluation protocol for VSL that is more sensitive to the high false positives issues of current approaches. To ensure overfitting is not hidden by the evaluation protocol, we suggest to ***rule out early stopping*** from weakly-supervised VSL evaluation, and instead always evaluate models after training them to convergence (or a large number of iterations). Note that early stopping defeats the purpose of weakly-supervised VSL, as it requires a fully annotated evaluation subset for tracking performance. To assess false detection of non-existing sound sources, we extend the evaluation subsets of both Flickr-SoundNet [13] and VGG Sound Sources [4] to include samples without visible sound sources. We also propose to use standard metrics that can measure the balance between high localization accuracy and low false positive rates.

Extended Flickr-SoundNet/VGG-SS. We extended VGG-SS/Flickr-SoundNet by merging clips with no sounding objects to the original test sets. Specifically, we analyzed 1000 videos from VGG-Sound test set (and 250 from Flickr-SoundNet test set), and manually selected 5-second clips with non-visible sound sources. This resulted in 379/42 samples with no sounding sources for VGG-SS/Flickr-SoundNet, respectively. Beyond manually identifying negative pairs, we further generate negative samples by pairing the audio and visuals from different videos. We control the difficulty of these negatives by sampling 25% from videos containing sources of the same class (hard negatives), and 75% from different classes. We merge all negatives with the VGG-SS [3] and Flickr-SoundNet [13] test sets. Table 1 shows the statistics of the extended test sets. For analysis, we also split test samples according to the size of the sound sources, as measured by the image area (in pixels) occupied by the ground-truth bounding boxes.

Evaluation metrics. Localization maps are often evaluated by comparing them to a group of human annotations using consensus intersection over union (cIoU, denoted as u) [13]. Given a set of predictions with cIoUs $\mathcal{U} = \{u_i\}_{i=1}^{N}$,

Table 1: Statistics of weakly-supervised audio-visual source localization test sets.

	Small	**Medium**	**Large**	**Huge**	**Total Pos**	**Real Neg**	**Automated Easy Neg**	**Automated Hard Neg**	**Total** Neg	**Total**
Ground-truth size (pixels)	1-32^2	32^2-96^2	96^2-144^2	144^2-224^2	1-224^2	0	0	0	0	1-224^2
Flickr-SoundNet [13]	0	9	83	158	250	0	0	0	0	250
Extended Flickr-SoundNet	0	9	83	158	250	42	169	39	**250**	**500**
VGG-Sound Source [3]	134	1796	1726	1502	5158	0	0	0	0	5158
Extended VGG-SS	134	1796	1726	1502	5158	379	3594	1185	**5158**	**10316**

Table 2: Comparison results of LocAcc for BEST and LATEST checkpoints. * denotes the value reported in the LVS [3] and HardPos[14] papers.

Method	**Flickr-SoundNet**		**VGG-SS**	
	Early Stop	NO Early Stop	Early Stop	NO Early Stop
Attention10k [13]	42.26	34.16	18.50*/18.52	14.04
CoarsetoFine [11]	–	47.20	–	21.93
DMC [5]	55.60	52.80	23.90	22.63
DSOL [6]	74.00	72.91	29.91	26.87
LVS [3]	71.90*/71.60	19.60	34.40*/33.36	10.43
HardPos [14]	76.80*	–	34.60*	–
EZ-VSL [7]	79.60	66.40	34.28	31.58
EZ-VSL + OGL [7]	**83.94**	**72.80**	**38.85**	**37.86**

prior work [5,1,11,3,14,7] measures the localization accuracy (LocAcc) among all samples with visible sounding objects, where each prediction is considered to be correct if its cIoU is above the cIoU threshold γ. This metric is also referred to as "CIoU". To avoid overriding nomenclature, we prefer the term Localization Accuracy. The cIoU threshold is set at $\gamma = 0.5$ unless otherwise specified.

Beyond localization error on samples with visible sounding objects (identified with a flag $c_i = 1$), we also evaluate on samples with no visible sounding sources ($c_i = 0$). Thus, the model is required to predict whether the current video contains a visible source or not. This is accomplished by computing a confidence score d_i, which we define as the maximum value in the predicted audio-visual similarity map. True positives are then given 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 $\mathcal{FN} = \{i|c_i = 1, d_i \leq \delta\}$. These sets are used to compute the Average Precision (AP), and the maximum F1 (max-F1) score obtained by sweeping the confidence threshold ($i.e.$ $\max_\delta F1(\delta)$).

3 Experiments

Datasets. We evaluate the effectiveness of the proposed method on two datasets - Flickr SoundNet [13] and VGG Sound Sources [4]. Following commonly-used settings [3,14,7], we use subset of 144k pairs for training in both cases. We then test on the respective extended test sets described in Sec. 2 and Tab. 1.

Prior works and baselines. We benchmark several prior VSL methods. Specifically, we considered Attention 10k [13], CoarsetoFine [11], DMC [5], DSOL [6], LVS [3], HardPose [14] and EZ-VSL [7]. We used authors' implementations when available, or our own otherwise. For EZ-VSL [7], we consider two versions: with and without object guided localization (OGL). OGL computes an object prior that is merged with the audio-visual localization map for improved predictions.

Table 3: Comparison results of the proposed metrics (AP, max-F1, Precision).

Method	**Extended Flickr-SoundNet**			**Extended VGG-SS**		
	AP	max-F1	Precision	AP	max-F1	Precision
CoarsetoFine [11]	0.00	38.20	47.20	0.00	19.80	21.93
LVS [3]	9.80	17.90	19.60	5.15	9.90	10.43
Attention10k [13]	15.98	24.00	34.16	6.70	13.10	14.04
DMC [5]	25.56	41.80	52.80	11.53	20.30	22.63
DSOL [6]	38.32	49.40	72.91	16.84	25.60	26.87
EZ-VSL [7]	46.30	54.60	66.40	24.55	30.90	31.58
EZ-VSL + OGL [7]	**48.75**	**56.80**	**72.80**	**27.71**	**34.60**	**37.86**

Table 4: Comparison results of max-F1 score for false positives among hand selected negatives, easy negatives and hard negatives.

Method	Extended Flickr-SoundNet			Extended VGG-SS		
	Real Neg	Automated Easy Neg	Automated Hard Neg	Real Neg	Automated Easy Neg	Automated Hard Neg
LVS [3]	14.50	17.80	14.30	8.30	8.80	7.60
Attention10k [13]	27.10	28.10	27.00	9.10	7.40	7.90
CoarsetoFine [11]	36.90	39.40	35.80	22.10	19.80	18.90
DMC [5]	48.80	41.40	43.70	20.70	19.40	21.90
EZ-VSL [7]	**52.60**	**55.70**	**54.20**	**32.80**	**36.10**	**31.40**

Preventing overfitting. To demonstrate that current methods suffer from severe overfitting, we trained models with and without early stopping [13,5,1,11,3,14,7]. Table 2 shows the localization accuracy (LocAcc) of these models on two datasets: Flickr SoundNet and VGG Sound Sources. Despite the large training sets (144k audio-visual pairs in both datasets), early stopping is critical to obtain high LocAcc. This observation suggests that, due to overfitting, prior methods do not scale well (*i.e.*, they cannot take advantage of larger datasets).

Preventing false positives. Since prior works rely on LocAcc as the main evaluation metric, models are not penalized for high false positives rates. To address this issue, we evaluated on the proposed Extended Flickr and VGG-SS datasets (without early stopping). Table 3 shows that, with the exception of EZ-VSL, prior works achieve relatively low AP and max-F1 scores, as they struggle to avoid false positives without substantially increasing false negatives.

Negative type. Table 4 studies the max-F1 score among a set containing only a particular type of negatives. For each of the three negative subsets, we add a similar number of positives. Once again, EZ-VSL was shown to better balance positive and negative detections, regardless of the type of negatives.

4 Conclusion

In this work, we identify the critical issue with current weakly-supervised visual sound source localization methods in their poor ability to identify when no sound sources are visible (*i.e.*, negatives). Since current evaluation protocols allow for early stopping and always assume the presence of visible sound sources, they are not sensitive to the aforementioned issues, the reason why they remained relatively unknown. To fix these issues, we propose a new evaluation protocol for VSL. We extend current evaluation datasets to also include negative samples (*i.e.*, frames with no visible sound source). These challenges are also addressed in a new framework presented in `https://github.com/stoneMo/SLAVC`.

References

1. Afouras, T., Owens, A., Chung, J.S., Zisserman, A.: Self-supervised learning of audio-visual objects from video. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 208–224 (2020)
2. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 609–617 (2017)
3. 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)
4. 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)
5. 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)
6. Hu, D., Qian, R., Jiang, M., Tan, X., Wen, S., Ding, E., Lin, W., Dou, D.: Discriminative sounding objects localization via self-supervised audiovisual matching. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). pp. 10077–10087 (2020)
7. Mo, S., Morgado, P.: Localizing visual sounds the easy way. arXiv preprint arXiv:2203.09324 (2022)
8. 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)
9. 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)
10. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 801–816 (2016)
11. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 292–308 (2020)
12. de Sa, V.R.: Learning classification with unlabeled data. Advances in neural information processing systems pp. 112–112 (1994)
13. 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)
14. 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)