AVSBench: A Pixel-level Audio–Visual Segmentation Benchmark

⋆Jinxing Zhou[1,2], ⋆Jianyuan Wang[3], Jiayi Zhang[2,4], Weixuan Sun[2,3], Jing Zhang[3], Stan Birchfield[5], Dan Guo[1], Lingpeng Kong[6,7], ✉Meng Wang[1], and ✉Yiran Zhong[2,7]

[1]Hefei University of Technology, [2]SenseTime Research, [3]Australian National University, [4]Beihang University, [5]NVIDIA, [6]The University of Hong Kong, [7]Shanghai Artificial Intelligence Laboratory
https://github.com/OpenNLPLab/AVSBench

Abstract. We propose to explore a new problem called audio-visual segmentation (AVS), whose goal is to output a pixel-level map of the object(s) that produce sound at the time of the image frame. To facilitate this research, we construct the first audio-visual segmentation benchmark (AVSBench), providing pixel-wise annotations for the sounding objects in audible videos. We propose two settings to be studied with AVSBench: 1) semi-supervised AVS with a single sound source and 2) fully-supervised AVS with multiple sound sources. We report the results of our baseline framework and six methods from the relevant tasks on our benchmark. Experiments demonstrate that AVSBench is promising for building a bridge between the audio and pixel-wise visual semantics.

Keywords: Audio-visual segmentation, Benchmarking, AVSBench.

1 Introduction

A human can classify an object not only from its visual appearance but also from the sound it makes. For example, when we hear a dog bark or a siren wail, we know the sound is from a dog or ambulance, respectively. Such observations confirm that the audio and visual information complement each other.

To date, some researchers have investigated the audio-visual correspondence (AVC) [1,2,3] problem, which aims to determine whether an audio signal and a visual image describe the same scene. Others studied audio-visual event localization (AVEL) [11,13,22,24,25,26,18,19,7,31], which classifies the segments of a video into the pre-defined event labels. Additionally, audio-visual video parsing (AVVP) [21,23,12,27] divides a video into several events and classify them as audible, visible, or both. Due to a lack of pixel-level annotations, all these scenarios are restricted to the frame/temporal level, thus reducing the problem to that of audible image classification.

⋆Equal contribution. ✉ Corresponding author (zhongyiran@gmail.com)

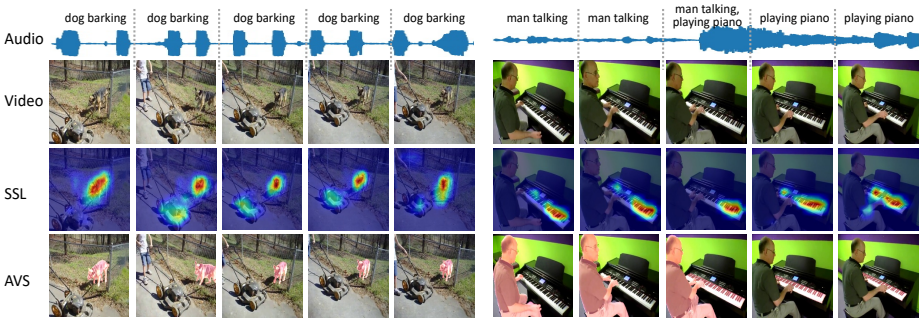

Fig. 1. Comparison of the proposed AVS task with the SSL task. Sound source localization (SSL) estimates a rough location of the sounding objects in the visual frame, at a patch level. We propose AVS to estimate pixel-wise segmentation masks for all the sounding objects, no matter the number of visible sounding objects.

A related problem, known as sound source localization (SSL), aims to locate the visual regions within the frames that correspond to the sound [1,2,6,4,10,17]. Compared to AVC/AVEL/AVVP, the problem of SSL seeks patch-level scene understanding, *i.e.*, the results are usually presented by a heat map that is obtained either by visualizing a similarity matrix or by class activation mapping (CAM) [29]. It does not consider the actual shape of the sounding objects.

Therefore, a pixel-level audio-visual segmentation (AVS) problem is desired to be explored. The goal of AVS is to densely predict whether each pixel corresponds to the given audio, *i.e.*, a mask of the sounding object(s) is required to generate. Fig. 1 illustrates the differences between AVS and SSL. The AVS task is more challenging than previous tasks as it requires the network to not only locate the audible frames but also delineate the shape of the sounding objects.

To facilitate the research, we propose AVSBench, the first pixel-level audio-visual segmentation benchmark that provides ground truth labels for sounding objects. We divide our AVSBench dataset into two subsets, depending on the number of sounding objects in the video (single- or multi-source). Correspondingly, there are two settings of audio-visual segmentation: 1) semi-supervised Single Sound Source Segmentation (S4), and 2) fully-supervised Multiple Sound Source Segmentation (MS3). We test six methods from related tasks on AVSBench and provide a new AVS method as a strong baseline. The extensive experiments verify the benefits of considering audio signals for visual segmentation, and the effectiveness of our proposed approach.

2 The AVSBench

Dataset Statistics. AVSBench is designed for pixel-level audio-visual segmentation. We collected the videos using the techniques introduced in VGGSound [5] to ensure that the audio and visual clips correspond to the intended semantics. AVSBench contains two subsets—Single-source and Multi-sources—depending on the number of sounding objects. All videos were collected from YouTube with

Fig. 2. AVSBench samples. The AVSBench dataset contains the Single-source subset (Left) and Multi-sources subset (Right). Each video is divided into 5 clips, as shown. Annotated clips are indicated by brown framing rectangles; the name of sounding objects is indicated by red text. Note that for Single-source training set, only the first frame of each video is annotated, whereas 5 frames are annotated for all other sets.

Table 1. Existing audio-visual dataset statistics. Each benchmark is shown with the number of videos and the *annotated* frames. The final column indicates whether the frames are labeled by category, bounding boxes, or pixel-level masks.

benchmark	videos	frames	classes	types	annotations
AVE [22]	4,143	41,430	28	video	category
LLP [21]	11,849	11,849	25	video	category
Flickr-SoundNet [20]	5,000	5,000	50	image	bbox
VGG-SS [4]	5,158	5,158	220	image	bbox
AVSBench (ours)	5,356	12,972	23	video	pixel

the *Creative Commons* license, and each video was trimmed to 5 seconds. The Single-source subset contains 4,932 videos over 23 categories, covering sounds from humans, animals, vehicles, and musical instruments. For the Multi-sources subset, we picked the videos that contain multiple sounding objects, *e.g.*, a video of baby laughing, man speaking, and then woman singing. To be specific, we randomly chose two or three category names from the Single-source subset as keywords to search for online videos, then manually filtered out videos to ensure 1) each video has multiple sound sources, 2) the sounding objects are visible, and 3) there is no deceptive sound, *e.g.*, canned laughter. In total, this process yielded 424 videos for the Multi-sources subset, out of more than six thousand candidates. The ratio of train/validation/test split percentages is set as 70/15/15 for both subsets. Several video examples are visualized in Fig. 2, where the red text indicates the name of sounding objects. As shown in Table 1, compared to other audio-visual datasets, our AVSBench contains 5,356 videos with 12,972 pixel-wise annotated frames, aiming to facilitate the research on fine-grained audio-visual segmentation.

Table 2. Comparison with methods from related tasks. Results of the evaluation metrics $\mathcal{J}$ and $\mathcal{F}$ under both S4 and MS3 settings are reported.

Metric	Setting	SSL		VOS		SOD		AVS (ours)	
		LVS[4]	MSSL[17]	3DC[14]	SST[8]	iGAN[15]	LGVT[28]	ResNet50	PVT-v2
$\mathcal{J}$	S4	.379	.449	.571	.663	.616	.749	.728	**.787**
	MS3	.295	.261	.369	.426	.429	.407	.479	**.540**
$\mathcal{F}$	S4	.510	.663	.759	.801	.778	.873	.848	**.879**
	MS3	.330	.363	.503	.572	.544	.593	.578	**.645**

Annotation. We divide each 5-second video into five equal 1-second clips, and we provide manual pixel-level annotations for the last frame of each clip. For this sampled frame, the ground truth label is a binary mask indicating the pixels of sounding objects, according to the audio at the corresponding time. For example, in the Multi-sources subset, even though a dancing person shows drastic movement spatially, it would not be labelled as long as no sound was made. In clips where objects do not make sound, the object should not be masked, *e.g.*, the *piano* in the first two clips of the last row of Fig. 2b. Similarly, when more than one object emits sound, all the emitting objects are annotated, *e.g.*, the guitar and ukulele in the first row in Fig. 2b. Also, when the sounding objects in the video are dynamically changing, the difficulty is further increased, *e.g.*, the second, third, and fourth rows in Fig. 2b.

Benchmark. We test the methods from related tasks (SSL, VOS, and SOD) on our benchmark. For each task, we pick two SOTA methods, and hence six in total. Additionally, we design a baseline for the AVS task and report its performance with two different backbones. The details of these six relevant methods and our baseline can be found in our main paper [30]. The quantitative results are shown in Table 2, with $\mathcal{J}$ measure [9] and $\mathcal{F}$ measure [16] as the evaluation metrics. The SSL methods show a substantial gap compared to our baseline, mainly because the SSL methods cannot provide pixel-level prediction. The SOTA SOD method LGVT [28] slightly outperforms our ResNet50-based baseline on the Single-source set ($\mathcal{J}$: 0.749 *vs.* 0.728), while is obviously worse than ours under the Multi-sources setting ($\mathcal{J}$: 0.407 *vs.* 0.479). This is because the SOD method relies on the dataset prior, and cannot handle the situations where sounding objects change but visual contents remain the same. Instead, the audio signals guide our method to identify which object to segment.

3 Conclusion

We have proposed a new task called AVS, which aims to generate pixel-level binary segmentation masks for sounding objects in audible videos. To facilitate research in this area, we collected the first audio-visual segmentation benchmark (called AVSBench). We proposed a baseline framework and compared it with several existing SOTA methods of the related tasks on AVSBench, and further demonstrated that our method can build a connection between the sound and the appearance of an object.

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 609–617 (2017)
2. Arandjelovic, R., Zisserman, A.: Objects that sound. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 435–451 (2018)
3. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. Advances in Neural Information Processing Systems **29** (2016)
4. 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)
5. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: VGGSound: A large-scale audio-visual dataset. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725 (2020)
6. Cheng, Y., Wang, R., Pan, Z., Feng, R., Zhang, Y.: Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In: Proceedings of the 28th ACM International Conference on Multimedia (ACM). pp. 3884–3892 (2020)
7. Duan, B., Tang, H., Wang, W., Zong, Z., Yang, G., Yan, Y.: Audio-visual event localization via recursive fusion by joint co-attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4013–4022 (2021)
8. Duke, B., Ahmed, A., Wolf, C., Aarabi, P., Taylor, G.W.: SSTVOS: Sparse spatiotemporal transformers for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5912–5921 (2021)
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV pp. 303–338 (2010)
10. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9248–9257 (2019)
11. Lin, Y.B., Li, Y.J., Wang, Y.C.F.: Dual-modality seq2seq network for audio-visual event localization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2002–2006. IEEE (2019)
12. Lin, Y.B., Tseng, H.Y., Lee, H.Y., Lin, Y.Y., Yang, M.H.: Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. Advances in Neural Information Processing Systems **34** (2021)
13. Lin, Y.B., Wang, Y.C.F.: Audiovisual transformer with instance attention for audio-visual event localization. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (2020)
14. Mahadevan, S., Athar, A., Ošep, A., Hennen, S., Leal-Taixé, L., Leibe, B.: Making a case for 3D convolutions for object segmentation in videos. arXiv preprint arXiv:2008.11516 (2020)
15. Mao, Y., Zhang, J., Wan, Z., Dai, Y., Li, A., Lv, Y., Tian, X., Fan, D.P., Barnes, N.: Transformer transforms salient object detection and camouflaged object detection. arXiv preprint arXiv:2104.10127 (2021)
16. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE transactions on pattern analysis and machine intelligence pp. 530–549 (2004)

17. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: Proceedings of the European conference on computer vision (ECCV). pp. 292–308. Springer (2020)
18. Ramaswamy, J.: What makes the sound?: A dual-modality interacting network for audio-visual event localization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4372–4376. IEEE (2020)
19. Ramaswamy, J., Das, S.: See the sound, hear the pixels. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2970–2979 (2020)
20. 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)
21. Tian, Y., Li, D., Xu, C.: Unified multisensory perception: Weakly-supervised audio-visual video parsing. In: Proceedings of the European conference on computer vision (ECCV). pp. 436–454. Springer (2020)
22. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 247–263 (2018)
23. Wu, Y., Yang, Y.: Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1326–1335 (2021)
24. Wu, Y., Zhu, L., Yan, Y., Yang, Y.: Dual attention matching for audio-visual event localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 6292–6300 (2019)
25. Xu, H., Zeng, R., Wu, Q., Tan, M., Gan, C.: Cross-modal relation-aware networks for audio-visual event localization. In: Proceedings of the 28th ACM International Conference on Multimedia (ACM). pp. 3893–3901 (2020)
26. Xuan, H., Zhang, Z., Chen, S., Yang, J., Yan, Y.: Cross-modal attention network for temporal inconsistent audio-visual event localization. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 279–286 (2020)
27. Yu, J., Cheng, Y., Zhao, R.W., Feng, R., Zhang, Y.: MM-Pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. arXiv preprint arXiv:2111.12374 (2021)
28. Zhang, J., Xie, J., Barnes, N., Li, P.: Learning generative vision transformer with energy-based latent space for saliency prediction. Advances in Neural Information Processing Systems (NeurIPS) **34** (2021)
29. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (2016)
30. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-visual segmentation. In: European Conference on Computer Vision (2022)
31. Zhou, J., Zheng, L., Zhong, Y., Hao, S., Wang, M.: Positive sample propagation along the audio-visual event line. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8436–8444 (2021)