

AutoSFX: Automatic Sound Effect Generation for Videos (Supplementary Material)

Anonymous Author(s)
Submission Id: 2603*

ABSTRACT

In this supplementary material, we provide: **i)** evaluations for the video segmentation ability of our *AutoSFX* and comparisons between it and SOTA audiovisual segmentation approaches; **ii)** failure cases of our *AutoSFX*.

ACM Reference Format:

Anonymous Author(s). 2018. AutoSFX: Automatic Sound Effect Generation for Videos (Supplementary Material). In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 AUTOSEFX V.S. AVS APPROACHES

In this section, we first introduce the experiment setup (§ 1.1) for audiovisual segmentation (AVS) task, including datasets, evaluation metrics, and implementation details. Then, we evaluate the performance of our *AutoSFX* in comparison to state-of-the-art methods for AVS (§ 1.2). We also provide mass segmentation results in this section.

1.1 Setup

Dataset. We leverage the AVSBench [10] dataset to evaluate the performance of our proposed approach. This is a recently released video segmentation dataset, providing masks for sounding objects with audio signals. It covers 23 object categories, e.g., animal and human-related sound events. Specifically, it consists of two subsets with different sound source protocols: S4, the Single-source subset, includes 4932 videos with 10,852 annotated frames; MS3, the Multi-source subset, includes 424 videos with 2,210 annotated frames. Fig. 1 demonstrates some examples of the dataset.

Evaluation Metrics. We quantify the performance by adopting standard segmentation and audio generation metrics outlined in [10]. Specifically, we quantify the segmentation performance using two benchmark settings: S4, where only a portion of the ground truth is provided during training, but all frames need to be predicted during evaluation; and MS3, where the labels of all five sampled frames of each video are available for training. With the evaluation metrics of mIoU (mean Intersection-over-Union), $\mathcal{M}_{\mathcal{G}}$, and F-score, $\mathcal{M}_{\mathcal{F}}$, we could measure the region similarity

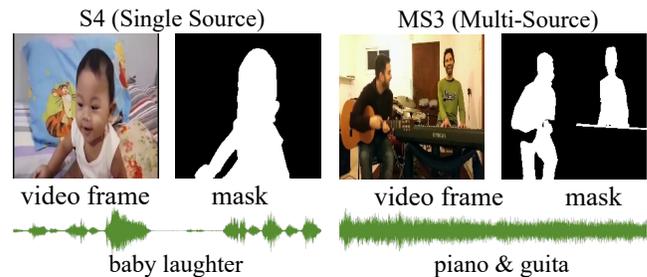


Figure 1: Examples of the AVSBench dataset used in our experiments.

and contour accuracy of the generated masks compared to that of ground truth.

Implementation Details. We resized all video frames to a size of 1024×1024 and extracted the log Mel-Spectrogram using 64 mel filter banks over 1 second of audio data sampled at 16,000 kHz on AVSBench. We employ Adam optimizer to optimize the model parameters with an initial learning rate of 10^{-4} with cosine decay. The batch size is defined as 8, and we train on S4 for 20 epochs and 40 epochs for MS3.

1.2 Quantitative Evaluation

In this section, we carry out a series of experiments to evaluate the performance of our proposed *Sound Generation Module* in terms of video segmentation, including the comparison with other SAM-based methods (§ 1.2.1), the comparison with methods from related tasks (§ 1.2.2), and model evaluation (§ 1.2.3).

1.2.1 Comparison with other SAM-based AVS methods. We conducted evaluations of various SAM-based methods on AVSBench, including SAM [3], AV-SAM [7], SAMA-AVS [4], and our *AutoSFX*. To establish a baseline, we utilize the released model weights of the ViT-H SAM model. This was done without any additional training. Furthermore, we adopted the training strategy proposed by SAMA-AVS to respectively train and test our model on the two subsets of AVSBench, i.e. S4 and MS3.

The quantitative results of the different SAM-based methods are demonstrated in Table 1. Overall, our *AutoSFX* has significant advantages over SAM, AVS, and AV-SAM on both subsets and both metrics, i.e. $\mathcal{M}_{\mathcal{G}} = .807$ and $\mathcal{M}_{\mathcal{F}} = .878$ for S4 and $\mathcal{M}_{\mathcal{G}} = .645$ and $\mathcal{M}_{\mathcal{F}} = .692$ for MS3. When focusing on the results from the S4 subset, we observe that our results surpass the performance of the vanilla SAM and AV-SAM. This may be due to our enhanced emphasis on auditory information in the video segmentation branch, leading to a more robust audio-visual fusion across the model rather than just in the initial stage. Moreover, the performance of our *AutoSFX* is comparable to that of SAMA-AVS, which further validates

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, June 03–05, 2018, Woodstock, NY
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
<https://doi.org/XXXXXXX.XXXXXXX>

Table 1: Comparison between different SAM-based methods on the test sets of S4 and MS3 of AVSBench. We demonstrate the results with the evaluation metrics of $\mathcal{M}_{\mathcal{J}}$ and $\mathcal{M}_{\mathcal{F}}$.

Metric	Setting	SAM [3]	AVS [10]	AV-SAM [7]	SAMA-AVS [4]	<i>AutoSFX</i> (ours)
$\mathcal{M}_{\mathcal{J}}$	S4	.551	.787	.408	.815	.807
	MS3	.540	.540	-	.631	.645
$\mathcal{M}_{\mathcal{F}}$	S4	.739	.879	.566	.886	.878
	MS3	.638	.645	-	.691	.692

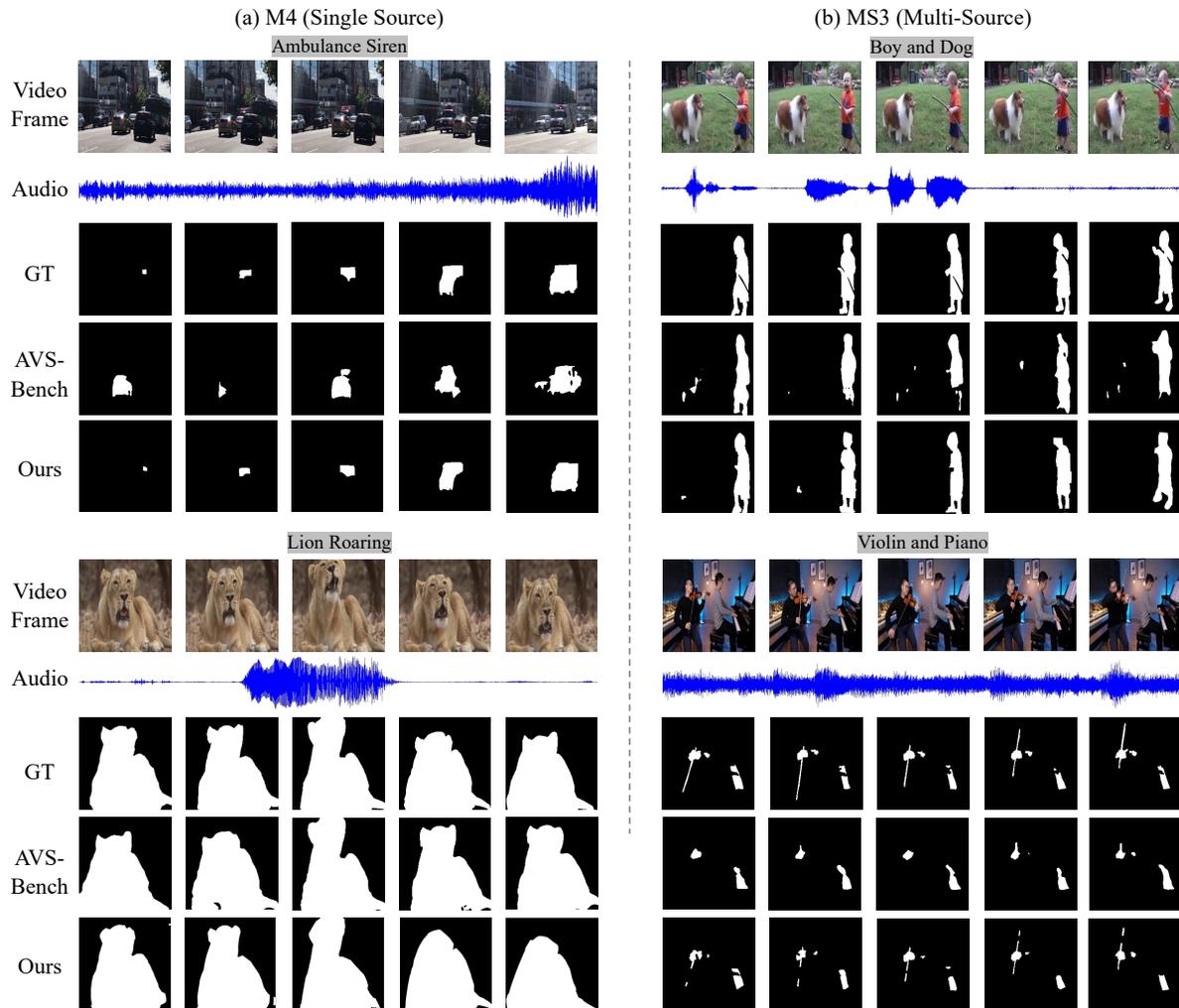


Figure 2: Qualitative examples of our proposed *BiGSAM*, under the semi-supervised M4 setting and fully-supervised MS3 setting, respectively.

our idea that the representations in the two branches (*i.e.* audio-visual segmentation and vision-to-sound generation) can mutually promote and enhance each other.

In Fig. 2, we demonstrate some qualitative results of the generated segmentation masks among our *AutoSFX*, AVSBench, and the corresponding ground truth. We observe that for both S4 and MS3

subsets, our results are closer to the ground truth. Specifically, on the test set of S4 (Fig. 2(a)), our method produces higher-quality results on single-sounding objects, *i.e.* clearer contour and finer details. Take the driving bus as an example, our method could effectively handle occlusion and achieve improved segmentation. Also, *AutoSFX* could precisely segment frames of a lion roaring, even

Table 2: Comparison with methods from related tasks on two subsets of AVSBench. We demonstrate the results with the evaluation metrics of $\mathcal{M}_{\mathcal{J}}$ and $\mathcal{M}_{\mathcal{F}}$.

Metric	Setting	SSL		VOS		SOD		<i>AutoSFX</i>
		LVS [1]	MSSL [8]	3DC [5]	SST [2]	iGAN [6]	LGVT [9]	(ours)
$\mathcal{M}_{\mathcal{J}}$	S4	.379	.449	.571	.663	.616	.749	.807
	MS3	.295	.261	.369	.426	.429	.407	.645
$\mathcal{M}_{\mathcal{F}}$	S4	.510	.663	.759	.801	.778	.873	.878
	MS3	.330	.363	.503	.572	.544	.593	.692

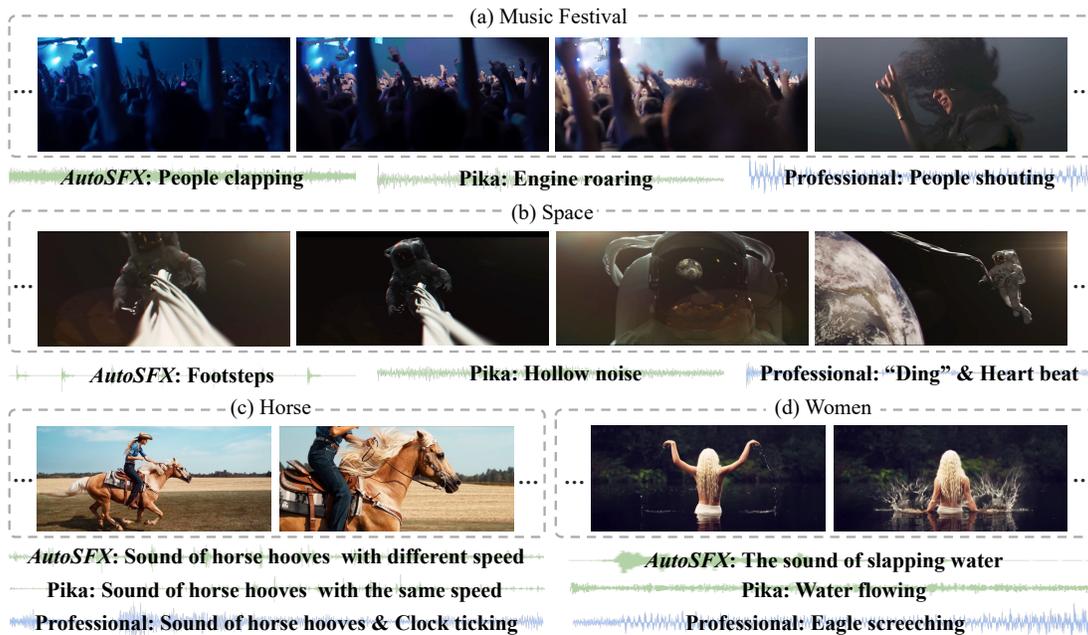


Figure 3: Failure cases of our *AutoSFX* and the online sound effect generation tool (Pika).

against a background with a similar texture. On the other hand, for the test set of MS3 (Fig. 2(b)), our generated masks have more precise boundaries for multiple-sounding objects, reducing ambiguous lines and spots, and overlapping areas.

1.2.2 Comparison with methods from related tasks. The most relevant tasks for AVS are sound source localization (SSL), video object segmentation (VOS), and salient object detection (SOD). We compare our framework with the methods from the three tasks. Specifically, we demonstrate two SOTA methods within each task, *i.e.* LVS [1] and MSSL [8] for SSL, 3DC [5] and SST [2] for VOS, and iGAN [6] and LVGT [9] for SOD. Note that the backbones of these methods are all pre-trained on the ImageNet. As shown in Table 2, the quantitative results demonstrate that our method achieves significantly superior segmentation performance than other methods.

1.2.3 Model Evaluation. Here we perform an ablation study to investigate the choice of different modules in our proposed *AutoSFX*.

Audio-Visual Fusion. We train the models with either cross-modal attention or easily concatenate the visual and auditory features

extracted from the corresponding encoders. For the latter strategy, we then obtained results of $\mathcal{M}_{\mathcal{J}} = .792$ and $\mathcal{M}_{\mathcal{F}} = .849$ for S4, only lower than that of SAMA-AVS [4], and $\mathcal{M}_{\mathcal{J}} = .607$ and $\mathcal{M}_{\mathcal{F}} = .675$ for MS3. On the contrary, leveraging our cross-modal attention mechanism, the model achieves a better performance.

Diffusion-based module. We perform the experiments by moving the diffusion-based module only for the audio generation branch, rather than for audio-visual diffusion. The results demonstrate that the model improves when the diffusion-based module considers multi-modal information. Specifically, the ablated framework achieves results of $\mathcal{M}_{\mathcal{J}} = .684$ and $\mathcal{M}_{\mathcal{F}} = .693$ for S4, and $\mathcal{M}_{\mathcal{J}} = .602$ and $\mathcal{M}_{\mathcal{F}} = .643$ for MS3.

2 FAILURES

As the performance of our *AutoSFX* is limited to the sound generation model trained on the VEGAS and VGGSound datasets, diverse user inputs pose experience failures in the sound effect generation. For example, as shown in Fig. 3 (a), generating sound effects for scenes featuring large crowds in dimly lit settings, *e.g.*, like many

people shaking hands at a music festival, can be challenging. Professional designers utilized shouts and cheers to create the complex soundscape, while the generated results are suboptimal; *i.e.* Pika's output resembles the roar of an engine, and ours merely produces the sound of clapping. There are some special scenarios in which sounds should be abstract, like in space (Fig. 3 (b)). In reality, space is silent, but designers would like to use heartbeat sounds to create a mysterious ambiance; *AutoSFX* and *Pika* generate very different results. Additionally, professional designers sometimes employ sound effects for visual content more abstractly, such as using the ticking of a clock to represent a running horse (Fig. 3 (c)) and an eagle's screech for a woman waving her arms like a bird. Future work could further incorporate text prompts during the training process of the generation module. We also believe there is potential for improvement in video understanding and generating abstract, indescribable, and context-aware sound effects.

REFERENCES

- [1] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2021. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16867–16876.
- [2] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. 2021. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5912–5921.
- [3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *ICCV*. 4015–4026.
- [4] Jinxiang Liu, Yu Wang, Chen Ju, Ya Zhang, and Weidi Xie. 2024. Annotation-free Audio-Visual Segmentation. In *WACV*.
- [5] Sabarinath Mahadevan, Ali Athar, Aljoša Ošep, Sebastian Hennen, Laura Leal-Taixé, and Bastian Leibe. 2020. Making a case for 3d convolutions for object segmentation in videos. In *British Machine Vision Conference 2020, BMVC 2020, Virtual Event*. BMVA Press.
- [6] Yuxin Mao, Jing Zhang, Zhexiong Wan, Yuchao Dai, Aixuan Li, Yunqiu Lv, Xinyu Tian, Deng-Ping Fan, and Nick Barnes. 2021. Transformer transforms salient object detection and camouflaged object detection. In *CoRR*.
- [7] Shentong Mo and Yapeng Tian. 2023. AV-SAM: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836* (2023).
- [8] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. 2020. Multiple sound sources localization from coarse to fine. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 292–308.
- [9] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. 2021. Learning generative vision transformer with energy-based latent space for saliency prediction. *Advances in Neural Information Processing Systems* 34 (2021), 15448–15463.
- [10] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. 2022. Audio-visual segmentation. In *ECCV*. Springer, 386–403.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009