THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Audio-Visual Scene Analysis with Self-Supervised Multisensory Features

Andrew Owens   Alexei A. Efros

UC Berkeley

Presented by: **Justin Chen, Yulu Pan, Liujie Zheng, Soumitri Chattopadhyay**

Dept. of Computer Science, UNC Chapel Hill

10/11/2023

Arguments

- Our pipeline is simple, intuitive and effective. PixelPlayer’s pipeline is way more complicated than ours.
- Their new MUSIC dataset only contains 685 videos
    - Unpopular dataset (101 stars on Github)
    - Only YouTube video IDs, what if the video gets deleted/corrupted?
- Their application is limited (only sound source localization and seperation) while ours has a wide range of applications in the audio-visual community
- They only test on the small MUSIC dataset, while ours test on more popular and large scale dataset. Ours has more quantitative results and more baselines.

The Sound of Pixels

Aniruddh Doki, Feihong He, Yue Yang, Ce Zhang

Department of Computer Science, University of North Carolina at Chapel Hill

Paper Battle

Oct 11st, 2023

Key Advantages of Our Paper (“The Sound of Pixels”) Over the Other One

- Point 1
  - For the “Sound localization” task: The output of [2] is a heat map that indicates whether a given pixel is likely (or unlikely) to be attributed to the audio. However, [2] **cannot distinguish which, of several, object instances is making a sound** (as shown in Figure 1).
  - But [1] could automatically show audios of several instances (as shown in Figure 2).

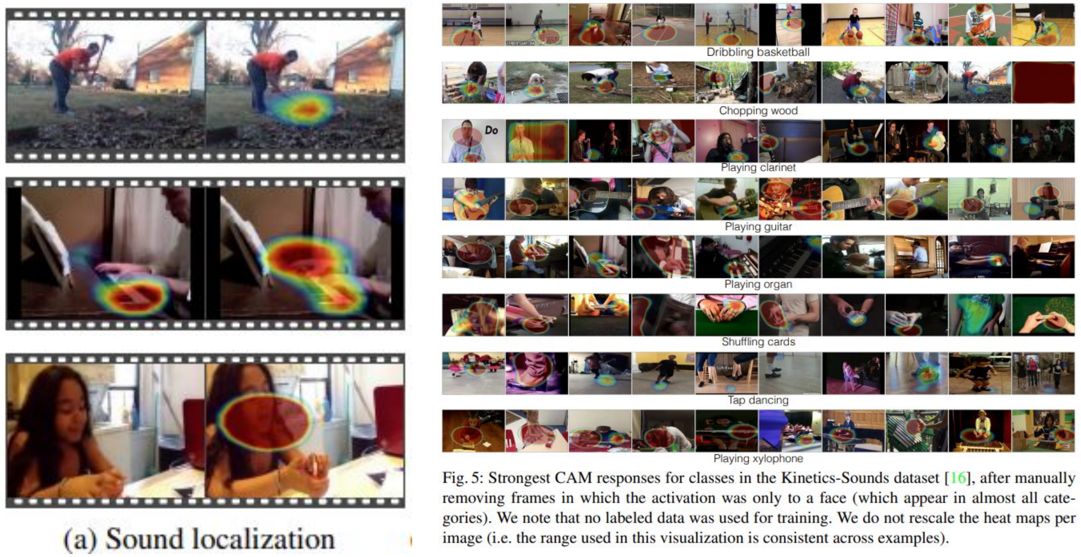


Fig. 5: Strongest CAM responses for classes in the Kinetics-Sounds dataset [16], after manually removing frames in which the activation was only to a face (which appear in almost all categories). We note that no labeled data was used for training. We do not rescale the heat maps per image (i.e. the range used in this visualization is consistent across examples).

Figure 1

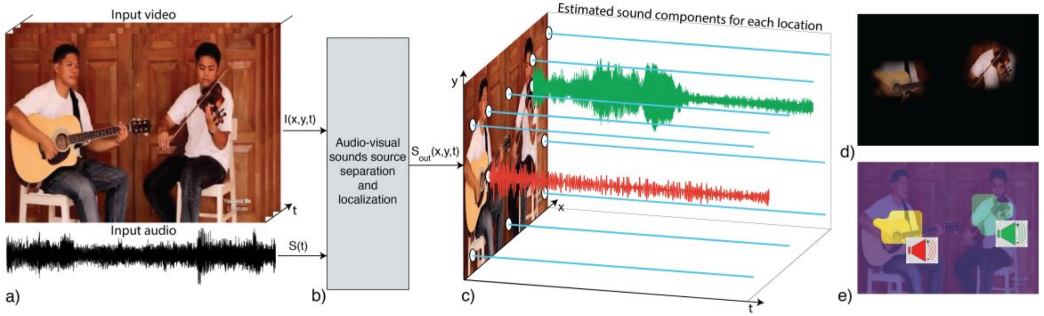
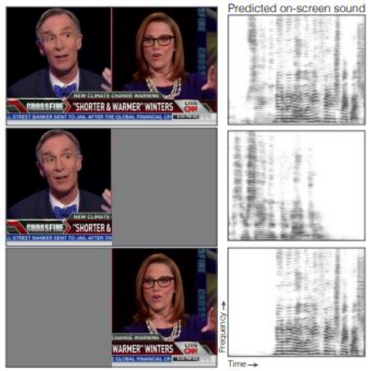

Input video
Input audio
a)
I(x,y,t)
S(t)
b)
Audio-visual sounds source separation and localization
S_out(x,y,t)
Estimated sound components for each location
y
x
t
c)
d)
e)

Figure 2

[1]. Zhao, Hang, et al. "The sound of pixels." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
[2]. Owens, Andrew, and Alexei A. Efros. "Audio-visual scene analysis with self-supervised multisensory features." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

Key Advantages of Our Paper (“The Sound of Pixels”) Over the Other One

- Point 2
  - Owing to the issue of [2] in Point 1, if you want to perform the “*audio separation*” task, you have to **manually** mask the corresponding part of the video.
    - Issue 1: Time-consuming because you need human involved.
    - Issue 2: How to mask itself (i.e., the mask size, mask shape, etc.) is already an issue. The example (Figure 1) given in the [2] is easy to mask, but what about harder examples (Figure 2, shown in paper [1])?
  - But [1] could do this audio in an end-to-end way.



(c) On/off-screen audio separation

Figure 1

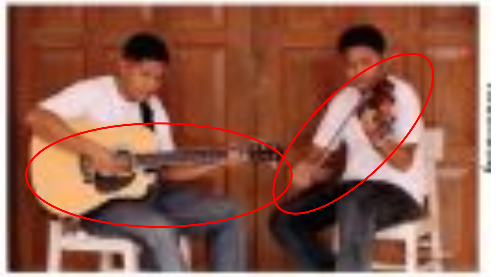


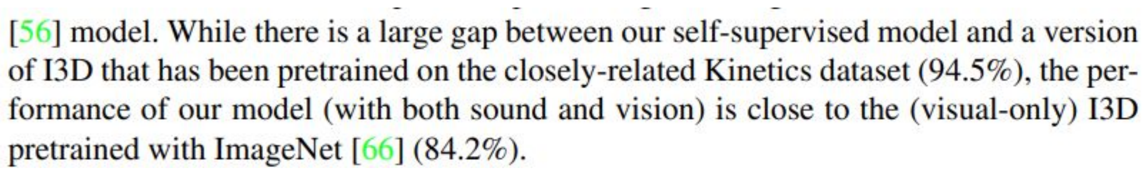
Figure 2. Two sources (red circles) of sound are very close

[1]. Zhao, Hang, et al. "The sound of pixels." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[2]. Owens, Andrew, and Alexei A. Efros. "Audio-visual scene analysis with self-supervised multisensory features." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

Key Advantages of Our Paper (“The Sound of Pixels”) Over the Other One

- Point 3
  - The way of [1] for self-supervised learning is based on combining several audios and separate them via the proposed method, so it’s targeted to the “*audio separation*” task. Therefore, although no comparison experiment is presented owing to different selections of dataset, I believe that [1] should have a better performance on this specific task.
- Point 4
  - [2] is applied to 3 tasks, which looks good. However, [1] already covers 2 of them (i.e., “*sound localization*” and “*audio separation*”). Most Importantly, [1] is an end-to-end method, but [2] needs extra finetune (even adding more NN layers!!!) for the 2 tasks.
  - As to the task (i.e., “*action recognition*”) that [1] doesn’t cover, we have 2 arguments.
    1. Because [1] targets on separating sound based on each pixel instead of trying to propose a self-supervised pre-training way, so it’s normal that [1] doesn’t work on this downstream task. Also, one thing I want to note is that the [1]’s NN architecture also contains those embeddings which could be useful for downstream task training. **Not doing this task doesn’t mean [1] cannot do well on it.**
    2. I think the “*action recognition*” task itself is a bit meaningless. Because image/video datasets are much larger than video+audio datasets, then when pre-training, why don’t I choose other pre-training methods on those larger datasets? Actually, [2] only achieves a similar performance as I3D on imagenet, much **worse** than I3D on Kinetics (as shown in the right-bottom screenshot).



[56] model. While there is a large gap between our self-supervised model and a version of I3D that has been pretrained on the closely-related Kinetics dataset (94.5%), the performance of our model (with both sound and vision) is close to the (visual-only) I3D pretrained with ImageNet [66] (84.2%).

[1]. Zhao, Hang, et al. "The sound of pixels." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[2]. Owens, Andrew, and Alexei A. Efros. "Audio-visual scene analysis with self-supervised multisensory features." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

Other battle parts

- Citation
  - Yes! [1] only has 489 citations, not as many as [2] (746).
  - **But** citation number cannot be an nice argument to say that [2] is better. [2] is a work targeting on pre-training and propose several results on 3 tasks, so it's easier to be followed by more groups because
    i. All groups working on video+audio will try [2]'s work for pre-training. [1] is an end-to-end method, so less likely to be tried by some groups.
    ii. All groups are related to the 3 tasks will cite [2]. But [1] doesn't explicitly propose the 2 tasks, which brings [1] less attention.
- Dataset
  - Yes! [2] use a more diverse dataset but [1] only use a music dataset.
  - **But**
    i. Nearly most of the good results from [2] is related to human talking, not very diverse.
    ii. We argue that music audio separation is an equally important problem as human speech separation. So using a more diverse dataset doesn't mean too much.

[1]. Zhao, Hang, et al. "The sound of pixels." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
[2]. Owens, Andrew, and Alexei A. Efros. "Audio-visual scene analysis with self-supervised multisensory features." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

Thank you!