Published in Image Processing On Line on 2024–05–29.
Submitted on 2024–01–26, accepted on 2024–05–13.
ISSN 2105–1232 © 2024 IPOL & the authors CC–BY–NC–SA
This article is available online with supplementary materials,
software, datasets and online demo at
https://doi.org/10.5201/ipol.2024.525

A Brief Analysis of SLAVC method for Sound Source Localization

Xavier Juanola, Gloria Haro

Universitat Pompeu Fabra
{xavier.juanola, gloria.haro}@upf.edu

Communicated by Quentin Bammey *Demo edited by* Aknine Billel and Xavier Juanola

Abstract

Mo and Morgado introduced in 2022 a novel self-supervised learning approach for Visual Sound Source Localization, denoted as SLAVC [Mo, S. and Mordado, P., A Closer Look at Weakly-Supervised Audio-Visual Source Localization, Advances in Neural Information Processing Systems, 2022]. The proposed method is based on multiple-instance contrastive learning. In addition to improving the results of previous methods, it also solves two critical problems that former methods faced: 1) excessive overfitting despite training on extensive datasets, 2) tendency to hallucinate sound sources even without visual evidence to support it in the video. In this paper, we briefly present the method, offer an online executable version allowing the users to test it on their own image-audio pairs and propose some improvements that could benefit the model as future work.

Source Code

The source code and documentation for this algorithm are available from the web page of this article[1]. The original implementation of the method is available here[2].
This is an MLBriefs article, the source code has not been reviewed!

Keywords: audio-visual; sound source localization

1 Introduction

In recent years, there has been a growing interest in exploring the correspondence between vision and hearing, particularly in the field of sound localization. The ability to locate the source of a sound is a fundamental aspect of human perception, enabling us to navigate our environment, communicate effectively, and respond to potential threats. Traditional research has focused extensively on visual

[1] https://doi.org/10.5201/ipol.2024.525
[2] https://github.com/stoneMo/SLAVC

cues, but recent investigations have highlighted the substantial impact of auditory information in this process. As a consequence, a new field known as visual sound source localization has emerged.

Visual sound source localization is the ability of humans to determine the spatial origin of a sound by integrating auditory and visual cues. By utilizing information from both senses, individuals are able to enhance their accuracy and precision in localizing sound sources, even in complex and dynamic environments. The integration between both worlds not only enriches our understanding of multisensory perception, but also holds potential for different applications, as virtual reality, robotics and assistive technologies for those who suffer from hearing impairment.

Early works that tried to solve the problem [7, 4, 10] learned the low-level correspondences between the hand-crafted visual features and the ones extracted from the audio. More recent approaches [15, 8, 1] use contrastive learning and localize objects aligning both visual and audio representation spaces.

Most of the previous methods presented two major flaws: 1) They needed to rely on early-stopping in order to avoid overfitting and 2) they were prone to visualize sound sources that were not in the scene because it was assumed that all sound sources were visible in the image. Thus, the model was identifying incorrectly (as false negatives) the sound sources inside the scene when those should be out of frame. On the other hand, some recent works started to focus on identifying silent objects and off-screen sounds [9, 12].

The work we are reviewing, [13], solves the problem of overfitting by applying a heavy visual dropout and momentum encoders, and addresses the problem of hallucinating visible sounding objects for off-screen sounds by using two terms: 1) a Visual Sound Localization term and 2) an Audio-Visual Correspondence term.

2 Method

The goal of Visual Sound Localization is to identify the regions in a video frame that contain the sources of the sounds present in its corresponding audio signal. As mentioned in the previous section, previous work to [13] suffered from two major problems: 1) overfitting, that occurred even when trained on large datasets and 2) hallucination of sounding objects in cases of off-screen sources, i.e., sounding objects not visible in the image (false negative identification). In contrast, the method reviewed in this paper – Simultaneous Localization and Audio-Visual Correspondence (SLAVC) [13] – addresses these two problems.

2.1 Getting Rid of Overfitting

To combat overfitting, Mo and Morgado used two regularization techniques: dropout [17] and momentum encoders [18, 5].

Dropout [17] is a regularization technique used in neural networks to prevent overfitting by randomly disabling a proportion of neurons during training, forcing the network to learn more robust and generalizable features. By randomly dropping out neurons, the network learns to rely on different combinations of neurons for each training example, effectively reducing the reliance on individual neurons and improving the network's ability to generalize to unseen data. This regularization technique is applied to the output of both visual and audio encoders. In order to prevent overfitting, a substantial visual dropout probability (as large as 0.9) had to be applied.

Momentum encoders [18, 5] are a specific type of neural network design that utilize a technique called momentum-based updates during training. By incorporating momentum, which involves considering past parameter updates, these encoders can effectively navigate intricate optimization landscapes and converge more quickly. By leveraging the direction and magnitude of previous gradients, momentum encoders efficiently learn complex representations, facilitating faster and more

stable convergence during the training process. They play a significant role in self-supervised and semi-supervised learning, as they enable the acquisition of slow-moving target representations. This leads to enhanced stability in self-training and improved quality of representations. Mo and Morgado apply momentum encoders to visual and audio inputs to obtain more stable targets.

2.2 Getting Rid of False Negatives

To achieve a better negative identification, Mo and Morgado propose a method that combines Visual Sound Localization and Audio-Visual Correspondence:

Simultaneous Localization and Audio-Visual Correspondence (SLAVC)

In order to perform Visual Sound Localization while reducing false positives in the presence of off-screen sounds, the authors propose the combination of two different terms:

- A localization term, P^{loc}, which discerns between the possible regions where the sound comes from and the unlikely regions. This term is the one in charge of the Sound Source Localization.
- An Audio-Visual Correspondence term, P^{avc}, that highlights regions in the data that are strongly correlated with the corresponding audio, while dampening regions that are better suited to be explained by other sound sources. When no sound sources are visible in the scene, this term creates ambiguity in the first term.

The two regularization methods explained in the previous section plus the Visual Sound Localization term are illustrated in Figure 1.

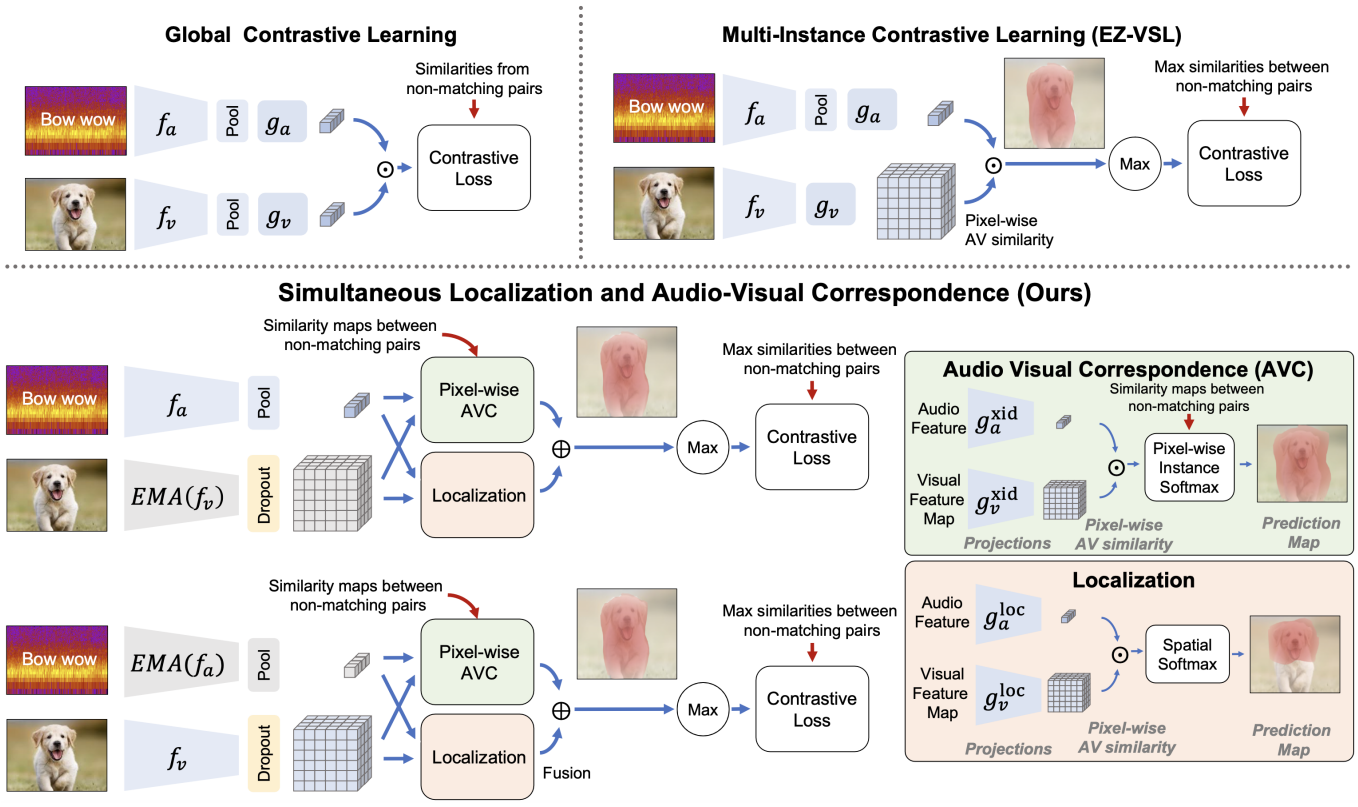


Figure 1: Visual Representation of the SLAVC method [13].

Then, the final prediction for the location where the sound source comes from takes into consideration both P^{loc} and P^{avc} and only selects regions that are simultaneously highlighted in both terms.

The two terms are based on different embedding spaces, one encoding the spatial localization and another the audio-visual correspondence. Let us denote an audio-visual dataset as $D = \{(\mathbf{v}_i, \mathbf{a}_i) : i = 1, \ldots, N\}$, where $\mathbf{a}_i$ and $\mathbf{v}_i$ are, respectively, the sound and the visual frame of the i-th instance. The audio representations in these two spaces are denoted as $g_a^{loc}(\mathbf{a}_i)$ and $g_a^{avc}(\mathbf{a}_i)$. Analogously, let $\{g_v^{loc}(\mathbf{v}_i^{xy}) : \forall x, y\}$ and $\{g_v^{avc}(\mathbf{v}_i^{xy}) : \forall x, y\}$, be the visual features, associated to locations (x, y) in the image, in those same spaces.

The localization term is defined as the softmax $\rho_{xy}(\cdot)$ over the spatial dimensions x and y

$$P^{loc}(\mathbf{a}_i, \mathbf{v}_j^{xy}) = \rho_{xy}\left(\frac{1}{\tau} s\left(g_a^{loc}(\mathbf{a}_i), g_v^{loc}(\mathbf{v}_j^{xy})\right)\right), \tag{1}$$

where τ is a temperature hyperparameter and $s\left(g_a^{loc}(\mathbf{a}_i), g_v^{loc}(\mathbf{v}_j^{xy})\right)$ is the cosine similarity[3] between an audio feature and a visual feature in the previous equation. The Audio-Visual Correspondence term is defined as the softmax $\rho_i(\cdot)$ over instances i

$$P^{avc}(\mathbf{a}_i, \mathbf{v}_j^{xy}) = \rho_i\left(\frac{1}{\tau} s\left(g_a^{avc}(\mathbf{a}_i), g_v^{avc}(\mathbf{v}_j^{xy})\right)\right). \tag{2}$$

As mentioned before, both terms, (1) and (2), are combined into a single prediction map

$$P^{SLAVC}(\mathbf{a}_i, \mathbf{v}_j^{xy}) = P^{loc}(\mathbf{a}_i, \mathbf{v}_j^{xy}) \cdot P^{avc}(\mathbf{a}_i, \mathbf{v}_j^{xy}). \tag{3}$$

Then, the model is trained to minimize the following loss

$$\mathcal{L}_i^{SLAVC} = -\log \frac{max_{xy} P(\mathbf{a}_i, \hat{\mathbf{v}}^{xy})}{\sum_{j=1}^{B} max_{xy} P(\mathbf{a}_i, \hat{\mathbf{v}}_j^{xy})} - \log \frac{max_{xy} P(\hat{\mathbf{a}}_i, \mathbf{v}^{xy})}{\sum_{k=1}^{B} max_{xy} P(\hat{\mathbf{a}}_i, \mathbf{v}_k^{xy})}, \tag{4}$$

where $\hat{\mathbf{a}}_i$ and $\hat{\mathbf{v}}_j^{xy}$ are the audio and visual momentum features.

2.3 Object Guided Localization (OGL)

At inference time, the authors combine the audio-visual similarities that the model have learned with a localization map towards the objects that appear in the scene.

The authors use the Object Guided Localization scheme presented in a previous paper, used in the EZVSL method [14]. The input image is passed to a convolutional model, f_{obj}, pre-trained on ImageNet [3] which has the same architecture than the visual encoder used for the audio-visual localization. This convolutional model yields a feature map $\mathbf{o} = f_{obj}(\mathbf{v}) \in \mathbb{R}^{C \times H \times W}$. It does not receive any information coming from the audio, thus the prediction is only based on the image itself.

Hence, this term can be used to tune the localization prediction towards the objects that appear in the scene, regardless of whether the sound is coming from the objects or not.

The approach that the authors use to extract object-centric localization maps without additional training relies on the fact that the convolutional model was trained on an object-centric dataset, and the activations are stronger when evaluated on images that contain objects. With this, they define the object localization as follows

$$S_{OGL}^{xy} = \|\mathbf{o}^{xy}\|_1. \tag{5}$$

[3]The cosine similarity is a measure of similarity between two vectors that measures the cosine of the angle between them.

2.4 Visual Sound Localization (VSL)

During inference, both localization and audio-visual correspondence terms are used for defining the Simultaneous Localization and Audio-Visual Correspondence (SLAVC) map

$$S^{xy}_{SLAVC} = s\left(g^{loc}_a\left(\hat{\mathbf{a}}\right), g^{loc}_v\left(\hat{\mathbf{v}}^{xy}\right)\right) + s\left(g^{avc}_a\left(\hat{\mathbf{a}}\right), g^{avc}_v\left(\hat{\mathbf{v}}^{xy}\right)\right). \tag{6}$$

Then, the Visual Sound Localization (VSL) map is computed as the linear aggregation between the SLAVC map with the Object Guided Localization (OGL) map (Equation (5)):

$$S^{xy}_{VSL} = \alpha S^{xy}_{SLAVC} + (1-\alpha) S^{xy}_{OGL}, \tag{7}$$

where $\alpha \in [0, 1]$ is a balancing term that weights the contribution of the OGL map versus the SLAVC one.

3 Training Details

The model presented by Mo and Morgado was trained on VGG Sound Sources [2], and in particular a subset of 144k samples, as used by [14, 19, 16].

For the audio and visual encoders, they followed prior work and used ResNet-18 [6]. ImageNet [3] pre-trained weights [14, 19, 16] are used to initialize the visual encoder.

In order to train the model, the authors used the Adam [11] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate of 10^{-4} and weight decay of 10^{-4}.

4 Demo

The online demo takes as input an image and an audio excerpt. The goal is to execute the SLAVC method and localize the visual sound sources in the image. The user is required to select the α value, which is the level of importance of SLAVC with respect of OGL in VSL (Equation (7)).

The demo outputs three images. All the images are resized to 224×224×3 as this is the size that the network works with. The three output images are: 1) Simultaneous Localization and Audio-Visual Correspondence (SLAVC) (Equation (6)), 2) Object Guided Localization (OGL) and 3) Visual Sound Localization (VSL) (Equation (7)). Those three images consist of a heatmap that overlays the input image.

5 Experiments

In order to do a qualitative analysis of the method proposed by Mo and Morgado, this section is organized as follows: 1) Analysis of the impact of the parameter α that appears in Equation (7) in the VSL output image, 2) examples of easy cases that the model is able to solve, 3) examples of some difficult cases that can not be solved by most of the existing methods at the time that paper [13] was published and 4) benefits of using OGL in the prediction of the localization map.

In order to assess the quality of the results, we will study images from the VGGSound dataset [2], as well as synthetic images that encompass the four challenging scenarios that will be mentioned in Section 5.3. This comprehensive approach will enable us to evaluate the performance across a range of difficult situations.

All the examples that will be shown in the following sections are available to be executed on the IPOL demo website, so one can run the model with the different images and audio excerpts reported

here. The default value of α is set to 0.4, the same value taken by Mo and Morgado in [14] for their analysis.

5.1 Impact of α

In this subsection we are going to analize the impact that α has in the output image corresponding to the VSL (weighted sum of SLAVC and OGL terms in Equation (7)).

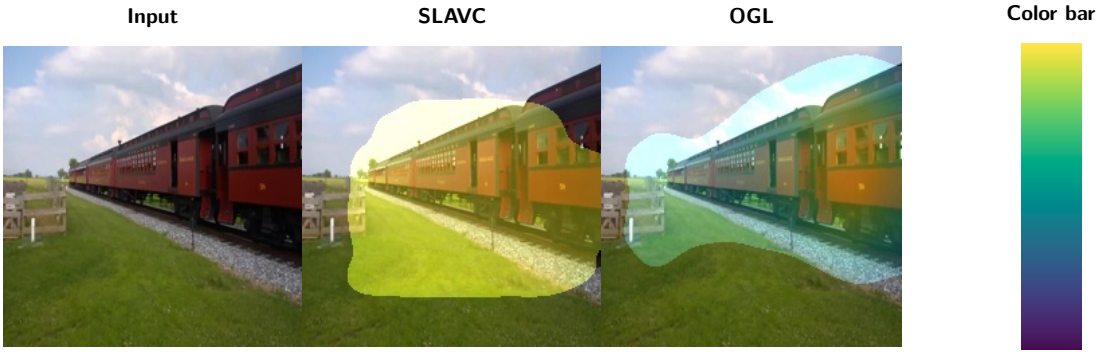


Figure 2: Input image, SLAVC and OGL of a train. (Colorbar at the right to understand color the scheme: brighter refers to a higher probability in the prediction).

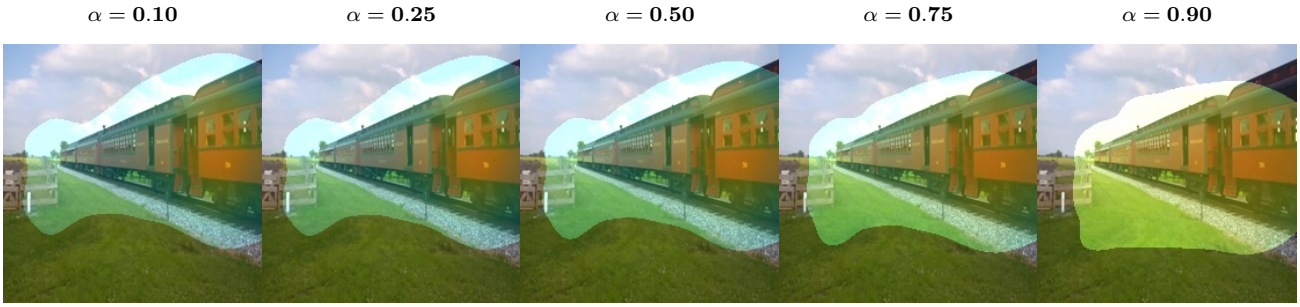


Figure 3: VSL executions for different values of α.

As expected, and as we can see from Figures 2 and 3, the smaller the value of α, the closer to the pure prediction of the OGL term, while the bigger the value, the closer to the SLAVC term, as expected by Equation (7).

5.2 Easy Cases

Figure 4 displays a grid consisting of four rows and four columns. The left column represents the input image provided to the model, while the subsequent three columns showcase the output results obtained by running the model on both the input image and an accompanying audio input. Each row corresponds to a distinct example, facilitating an in-depth analysis of the model's behaviour in various scenarios and with different types of sound sources.

The first row shows the back of a bird, the model accurately localizes the bird in the image, based on both the input audio of a talking parrot and the input image. The second image is a pair of cats, and the input audio are both cats purring. The model correctly highlights both cats. The third row presents an intriguing example — a composition of two views of the same guitar — accompanied by an audio of the guitar being played. The model successfully highlights different areas, including the fingers and the image on the top left, which is expected to be the sound source. Hence, the model accurately localizes the source of the sound. Lastly, the final example displays two vacuum cleaners,

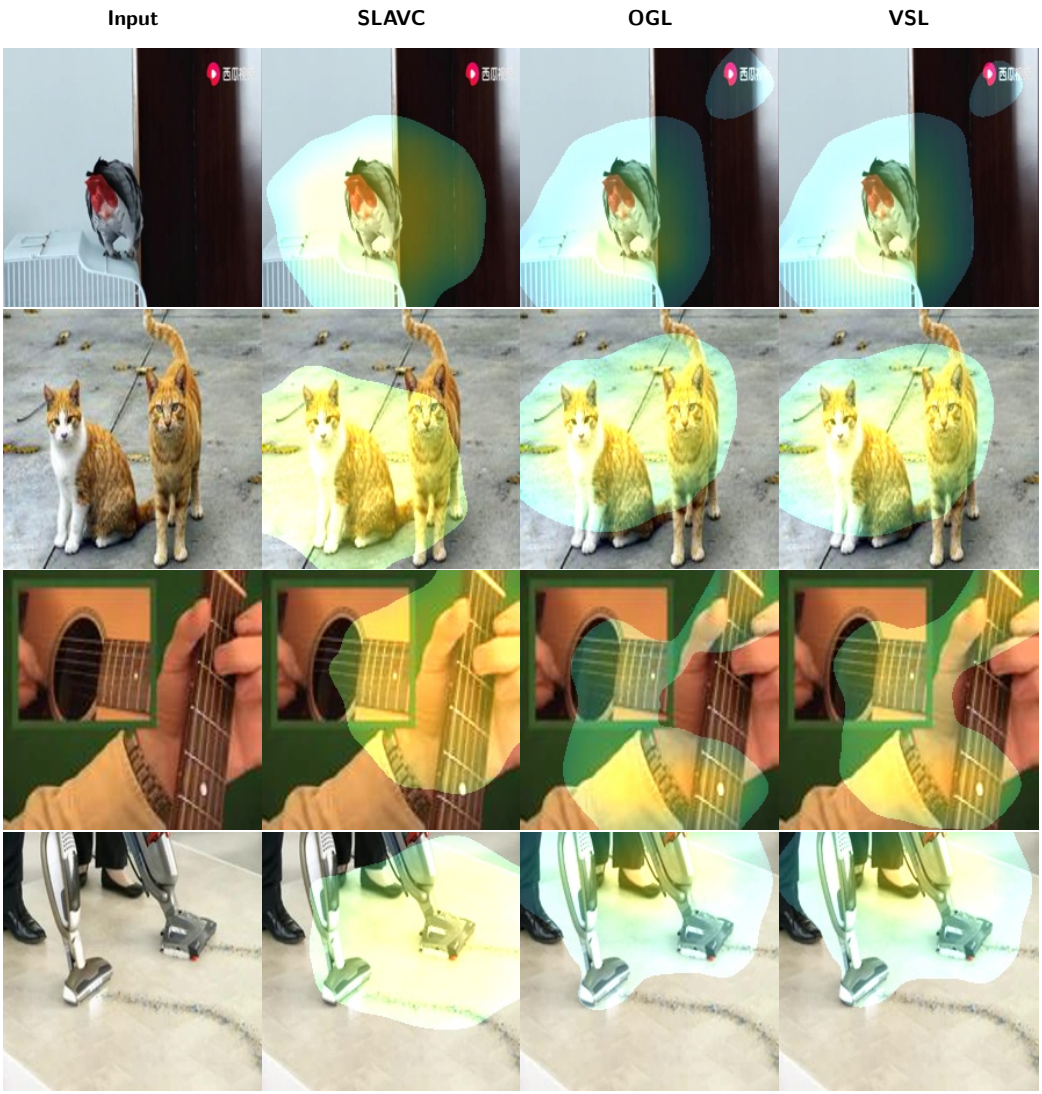


Figure 4: Easy examples that the model is able solve correctly.

with the audio input representing the sound produced by the machines. Here, the model correctly highlights the area from which the sound is being emitted.

Overall, the model demonstrates accurate sound source localization across various challenging scenarios, as evidenced by these examples.

5.3 Difficulties in Visual Sound Localization

We have identified some difficult cases in which the model is not able to correctly solve the Visual Sound Localization problem. Those cases are the following:

- In scenarios where the scene contains a mixture of sounds, with different audio sources present in the video, it is possible for certain sounds to be masked by others due to variations in volume. This can pose challenges to the localization process. Moreover, most of the existing visual sound localization models have been designed for the single-source case.

- When small objects within a scene emit sounds, their detection and localization can pose challenges, particularly if these objects are significantly smaller in size compared to the surrounding elements.

- When there are soundless objects present in a scene, algorithms often mislabel them as sound-emitting objects.

- Algorithms encounter difficulties in identifying sounds produced by objects that are not visible in the scene, either because they are occluded or directly out of frame (off-screen sounds).

- When multiple instances of the same type are present in the scene, algorithms suffer to identify which of them are producing the sounds. Moreover, if some of these instances are silent, the difficulty increases.

5.3.1 Mixture of Sounds

In Figure 5, two images and their corresponding results after running the SLAVC model are presented. In the first image, we observe two people playing ping pong. The audio corresponding to this image includes the voices of the two people talking, the sound of the ball bouncing on the table, and the footsteps of the players. Examining the resulting images, we can see that the highlighted regions include the ping pong table, where the ball is bouncing, as well as the two people playing. However, none of the resulting images highlight the feet. The feet should also be included in the VSL result, as the audio includes the sound of the sneakers slipping on the floor.

Moving on to the second image, it depicts an urban scene where there are cars on the different streets and people walking. The corresponding audio includes people talking, the sound of tires drifting, and the car engines. Upon analyzing the resulting images, we observe that the sound sources related to the cars (tires drifting and car engine) are correctly highlighted in the images, but the people talking are not correctly covered. Therefore, the model doesn't successfully solve this case.

Consequently, it is evident that the model is not capable of correctly locating sounds in scenarios where there is a mixture of different sound sources.

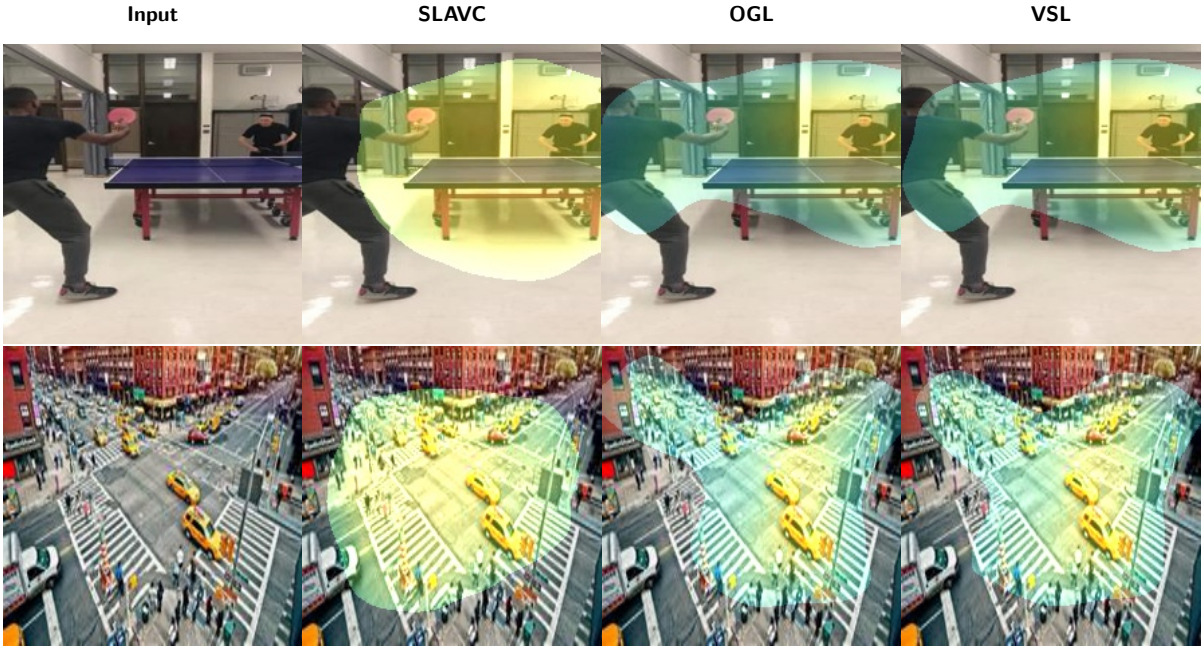


Figure 5: Examples of images where there are multiple sound sources present in the scene. In the first example (top row) there is people playing ping pong and there is a mixture of sounds between the people talking, the ball bouncing and the steps. The second example (bottom row) is an urban scene where we can hear the cars engines and people talking.

5.3.2 Small Objects

Figure 6 presents two input images and the results obtained after applying the SLAVC method to localize sound sources. In the first column of the first row, an image of a road is displayed, with a red car positioned at the right lane near the end of the road. The SLAVC method is executed with an audio file of a car. Upon examining the results, we observe that the SLAVC primarily highlights the road, indicating that the sound originates from that area. However, it fails to accurately localize the car within the image. The OGL term highlights various regions unrelated to the car. Finally, the VSL approach mainly selects the road. Consequently, the localization in this example is not correct.

Moving to the second row, an image of the sky is presented, featuring a tree on the left side and a small fly positioned near the center of the image. When feeding the model with this image and an audio file corresponding to a fly, we observe that all three images (SLAVC, OGL, and VSL) highlight the portion of the tree capturing only the fly at the border of the highlighted zone, meaning that the predicted sound source is far from the actual fly.

Based on these examples, it can be concluded that the model encounters difficulties when localizing small objects accurately.

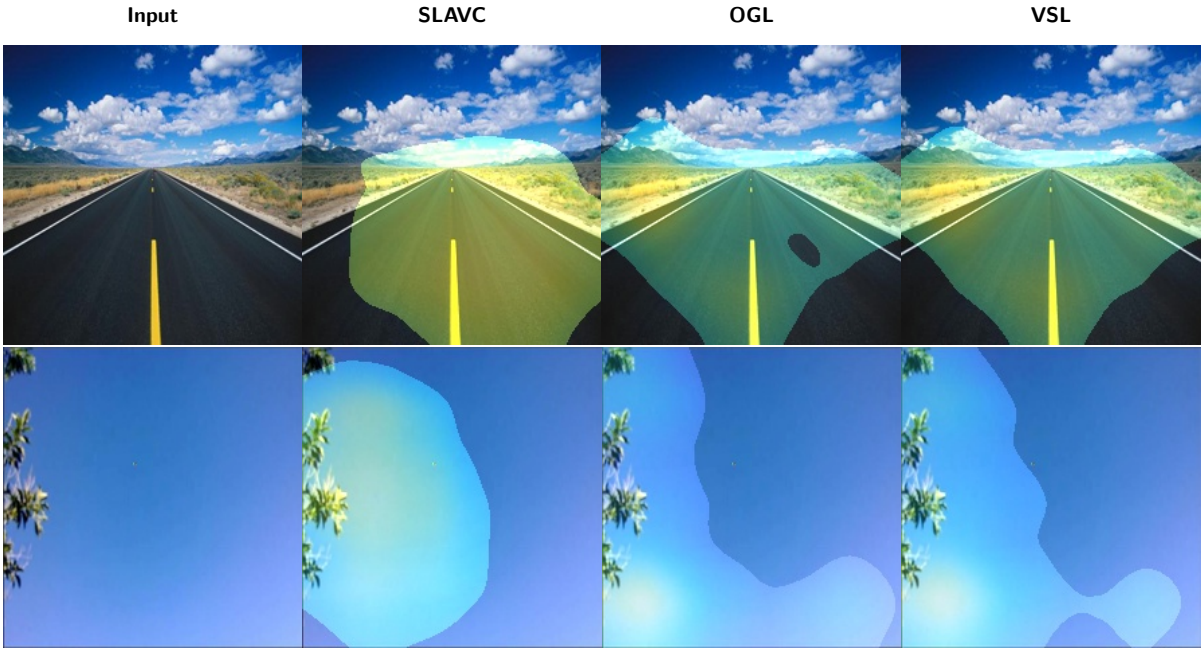


Figure 6: Examples of small objects and the corresponding results for the input images after applying the SLAVC method to localize sound sources. In the first row, there is a car very small at the end of the road. In the second row, there is a very small fly in the middle of the image.

5.3.3 Silent Objects

Figure 7 depicts two input images and their corresponding results, similar in structure to Figure 6. In the first row, the images feature two vacuum cleaners operated by different individuals. The sound input provided to the model consists of people talking and a child laughing. We observe that the SLAVC, OGL, and VSL images highlight the vacuum cleaners, leading to an incorrect localization of the sound. In this case, the objects (vacuum cleaners) should ideally be silent.

Likewise, in the second row, the image portrays three individuals playing billiards, with a child in the arms of the rightmost person. The sound input given to the model is that of a mosquito flying. It is noteworthy that only the SLAVC image highlights all the scene, including the people, while the OGL image highlights a smaller region of the scene and subsequently the VSL image highlights

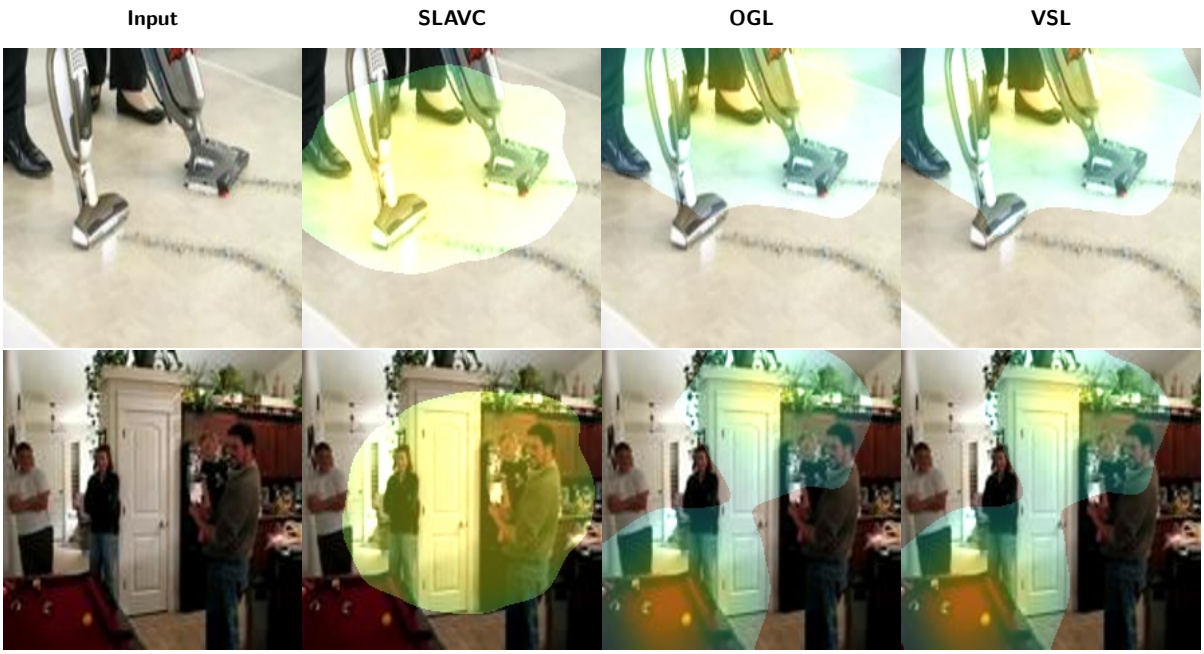


Figure 7: Examples of images where the primary objects that appear in the scene are silent. In the first row, there are two vacuum cleaners driven by people. In the second row, there is people.

areas where people are not located. In this instance, the model successfully identifies that the sound does not originate from the people but from the space between them, where the mosquito could potentially fly. Additionally, the model incorrectly highlights the billiards board.

Regarding the localization of silent objects, the model encounters some challenges but demonstrates the ability to handle such cases to a certain extent.

5.3.4 Off-screen Sounds

In Figure 8, two examples are presented where the sound sources are located outside of the scene. In the first row, the image showcases a forest with visible branches but no visible birds. The corresponding audio input passed to the model is that of birds singing. Upon examining the results for this example, we observe that the SLAVC image primarily highlights the trees located in the center of the scene, while the OGL image highlights regions slightly more towards the left. The model appears to localize sound in those areas; however, the sound sources (birds) are actually outside of the frame and thus not visible.

In the second example, we have one of the images from Figure 7, but in this case we are passing the audio of a firefighter's siren that could be passing near the house. In this case, the locations highlighted are the same as in Figure 7, indicating that the model detects similar correlations between the image and audio for different audio inputs. It seems that when there is no sound source within the frame, the model's decision is predominantly based on the image alone.

Consequently, it is evident that the model encounters difficulties in recognizing and correctly highlighting regions when the sound sources are not present in the image.

5.3.5 Different Objects of the Same Type

In Figure 9, three different images are presented, each showing three different instances of the same type of object. The first image displays an orchestra with four people playing the tuba, the second image depicts three flying planes, and the third image shows three guitarists playing the guitar.

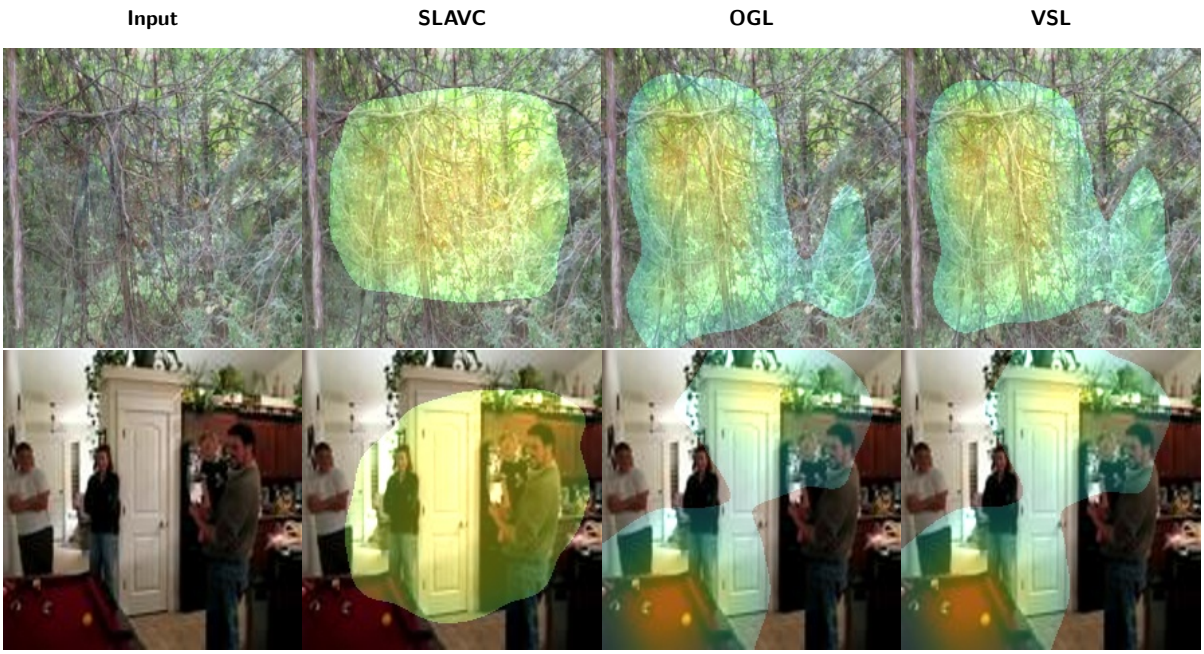


Figure 8: Examples of images where the sound source is located out of the screen. In the first row we see an image of a forest and its corresponding audio are birds singing. The second image is an image of people playing billiards with a kid, and the corresponding sound is a firefighter's siren.

In the three images and their corresponding results, we can see that the highlighted regions comprise all the objects (tubas in row 1, planes in row 2 and guitars in row 3). If we only have a still image and an audio, in which different objects of the same type are sounding (people talking, instruments that are played, or vehicles moving) in most of the cases, both an algorithm and a human will have trouble discerning which is the sound source.

To accurately determine the location of a sound source, additional contextual information is often required. Motion cues, such as the movement of the different objects in the scene can certainly help in sound localization.

5.4 SLAVC vs. VSL

In this subsection, we compared the predictions from the SLAVC method (Equation (6)) and the ones from the VSL (Equation (7)) which takes into consideration the SLAVC and the OGL.

A similar ablation was done in the previous paper published by the authors (EZVSL [14]). They reported that only using the OGL already surpassed prior state-of-the-art due to the fact that: 1) it is likely that the sound is originated at a location where an object lays and 2) the majority of test samples in VGGSound Sources and Flickr only contain a single sounding object in the scene. Moreover, the combination of the localization term with the OGL improved their results with respect to only using either the localization term or the OGL term.

In some cases in which multiple objects are in the scene, or the sound comes from the object that is mainly visible in the scene (like most of the images in the mentioned datasets which have a single source present in the scene), the OGL better adjusts the localization map from the VSL towards the objects. We can see this behaviour in the last example of Figure 4 and in all the examples where different objects of the same type appear (Figure 9). In these cases, the OGL term refines the VSL prediction towards the position of the objects present in the image.

On the other hand, this term worsens the prediction of the SLAVC in the cases where the salient objects in the images are not the ones producing the sound. It is interesting to look at the first example from Figure 4, in which the SLAVC correctly localizes the sound coming from the bird,

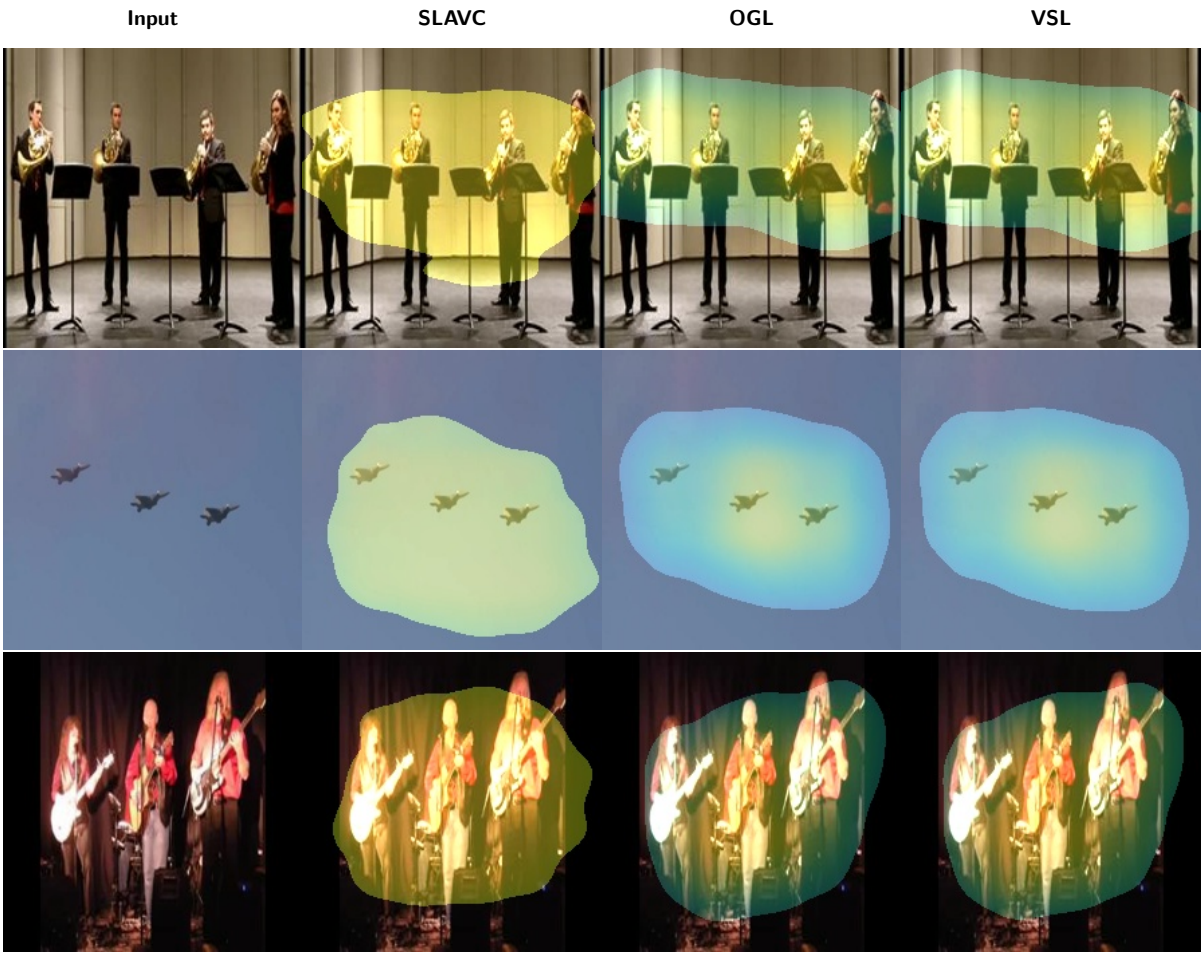


Figure 9: Examples of images where there are multiple objects of the same type in the image. The corresponding sounds are multiple objects of the same type sounding at the same time. In the first image we see an orchestra with different Tuba. In the second image there are three planes flying and in the corresponding sound we can hear the planes. Lastly, the third image is a band in which we can see and hear three guitars and people singing.

but the OGL also highlights the logo of the TV show located at the top right corner of the scene as an object, and directs the prediction towards it. Another interesting example where the OGL distracts the prediction is in Figure 6, specifically in the example in the second row. As we mentioned previously, there is a small fly in the image, but the model highlights the tree. This tree is completely highlighted in the OGL prediction, contrary to the SLAVC prediction. In this case, instead of helping the prediction to correctly highlight the sounding source, it is distracting the model and is making the prediction worse.

The same happens for the other difficult cases, like the silent objects (Figure 7) and off-screen sounds (Figure 8).

6 Conclusions and Proposed Improvements

In this paper, we have analyzed the SLAVC method presented in [13]. We have discussed the results obtained from the method through qualitative analysis in various challenging scenarios, including small objects emitting sound, silent objects, off-screen sounds, mixtures of sounds, and different instances of the same type of sound source.

Overall, SLAVC achieves really good results in many cases, however, there are still limitations in more general scenarios, as highlighted in our analysis.

To improve the results discussed in the reviewed paper, several improvements can be considered. Firstly, generalizing the input from static images to videos would enable the network to learn not only semantic information from the scene but also motion cues. By incorporating motion information, the model could better handle situations where multiple objects of the same type are present.

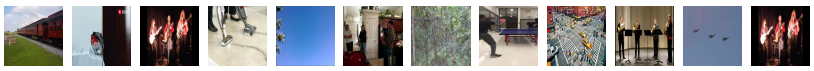
Additionally, the robustness to both silent objects and off-screen sounds could be improved following the ideas presented in [12], where audio and visual prototypes are used to define proper filters to be applied in the localization map and avoid errors in those difficult cases.

With these improvements, most of the problems proposed in this review would be addressed, thus improving the results of the method and being able to localize the sound sources in a more general scenario.

Acknowledgment

This work has been supported by MICINN/FEDER UE project PID2021-127643NB-I00.

Image Credits

 VGGSound dataset [2].

References

[1] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, *Localizing Visual Sounds the Hard Way*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16867–16876, https://doi.org/10.1109/CVPR46437.2021.01659.

[2] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, *VGGSound: A Large-Scale Audio-Visual Dataset*, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 721–725, https://doi.org/10.1109/ICASSP40776.2020.9053174.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *ImageNet: A Large-Scale Hierarchical Image Database*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

[4] J. W. Fisher III, T. Darrell, W. Freeman, and P. Viola, *Learning Joint Statistical Models for Audio-Visual Fusion and Segregation*, in Advances in Neural Information Processing Systems, vol. 13, 2000. https://proceedings.neurips.cc/paper_files/paper/2000/file/11f524c3fbfeeca4aa916edcb6b6392e-Paper.pdf.

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, *Momentum Contrast for Unsupervised Visual Representation Learning*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9729–9738, https://doi.org/10.1109/CVPR42600.2020.00975.

[6] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[7] J. Hershey and J. Movellan, *Audio Vision: Using Audio-Visual Synchrony to Locate Sounds*, in Advances in Neural Information Processing Systems, vol. 12, 1999. https://proceedings.neurips.cc/paper_files/paper/1999/file/b618c3210e934362ac261db280128c22-Paper.pdf.

[8] D. Hu, F. Nie, and X. Li, *Deep Multimodal Clustering for Unsupervised Audiovisual Learning*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9248–9257, https://doi.org/10.1109/CVPR.2019.00947.

[9] D. Hu, R. Qian, M. Jiang, X. Tan, S. Wen, E. Ding, W. Lin, and D. Dou, *Discriminative Sounding Objects Localization Via Self-Supervised Audiovisual Matching*, in Advances in Neural Information Processing Systems, vol. 33, 2020. https://proceedings.neurips.cc/paper_files/paper/2020/file/7288251b27c8f0e73f4d7f483b06a785-Paper.pdf.

[10] E. Kidron, Y. Y. Schechner, and M. Elad, *Pixels that Sound*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 88–95, https://doi.org/10.1109/CVPR.2005.274.

[11] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, ArXiv Preprint ArXiv:1412.6980, (2014), https://doi.org/10.48550/arXiv.1412.6980.

[12] X. Liu, R. Qian, H. Zhou, D. Hu, W. Lin, Z. Liu, B. Zhou, and X. Zhou, *Visual Sound Localization in the Wild by Cross-Modal Interference Erasing*, in AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 1801–1809, https://doi.org/10.1609/aaai.v36i2.20073.

[13] S. Mo and P. Morgado, *A Closer Look at Weakly-Supervised Audio-Visual Source Localization*, in Advances in Neural Information Processing Systems, vol. 35, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/file/f3f2ff9579ba6deeb89caa2fe1f0b99c-Paper-Conference.pdf.

[14] ——, *Localizing Visual Sounds the Easy Way*, in European Conference on Computer Vision (ECCV), 2022, pp. 218–234, https://doi.org/10.1007/978-3-031-19836-6_13.

[15] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, *Learning to Localize Sound Source in Visual Scenes*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4358–4366, https://doi.org/10.1109/CVPR.2018.00458.

[16] A. Senocak, H. Ryu, J. Kim, and I. S. Kweon, *Learning Sound Localization Better from Semantically Similar Samples*, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 4863–4867, https://doi.org/10.1109/ICASSP43922.2022.9747867.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *Dropout: a Simple Way to Prevent Neural Networks from Overfitting*, The Journal of Machine Learning Research, 15 (2014), pp. 1929–1958. https://dl.acm.org/doi/10.5555/2627435.2670313.

[18] A. Tarvainen and H. Valpola, *Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results*, Advances in Neural Information Processing Systems, 30 (2017). https://dl.acm.org/doi/10.5555/3294771.3294885.

[19] A. Vedaldi, H. Chen, W. Xie, T. Afouras, A. Nagrani, and A. Zisserman, *Localizing Visual Sounds the Hard Way*, Institute of Electrical and Electronics Engineers, 2021, https://doi.org/10.48550/arXiv.2104.02691.