# Learning from Silence and Noise for Visual Sound Source Localization

Xavier Juanola [1] Giovana Morais [2] Magdalena Fuentes [2] Gloria Haro [1]

[1]Universitat Pompeu Fabra [2]New York University

## Motivation

1. Most VSSL models perform poorly on **negative audio** (silence, noise, and *offscreen* sounds), where audio–image correspondence is weak.
2. Evaluation usually focuses only on **positive cases** and on datasets where sound sources are visible, centered, and in the foreground, which reduces the incentive to actually use the audio track for localization.

## Contributions

- **SSL-SaN model:** A self-supervised model trained with negative audio (Silence + Noise) that suppresses non-informative signals and improves localization and cross-modal retrieval.
- **IS3+ dataset:** A curated and extended version of IS3 with corrected labels, simplified classes, normalized audio at 16 kHz, and additional negative test cases (Silence, Noise, Offscreen). Each image is paired with two clean audio samples (from Adobe SFX and curated IS3 clips), with noisy or mismatched ones removed. IS3+ provides more reliable evaluation by avoiding the noisy audio issues of VGG-SS and IS3.
- **Separability metric:** A new metric that measures separation between positives and negatives, complements cIoU and AUC, and reflects real-world localization performance.

## Data and Evaluation

**Training**: VGGSound-144K (a fixed subset of VGGSound). **Testing**: VGG-SS (boxes), IS3, **IS3+** (our curated version with negatives), and AVS-Bench S4 (masks). We evaluate both **positives** and **negatives** (silence, noise, offscreen).

**IS3+ (curation)**: manual reassignment of incorrect labels, simplification of visually indistinguishable classes, replacement of audios with clean items from the same category, and normalization to 16 kHz.

## Learning from Silence and Noise

We make SSL-SaN robust to Silence and Noise:

- Each image is paired with silence ($a^S$) and Gaussian noise ($a^N$) which are used as additional negative audios for the contrastive learning.
- We add two loss terms enforcing empty similarity maps:

$$\mathcal{L}_S = \|S(\mathbf{a}^S, \mathbf{v}_j)\|_2^2, \quad \mathcal{L}_N = \|S(\mathbf{a}_j^N, \mathbf{v}_j)\|_2^2$$

These penalize activations under silence/noise and also improve performance with positive audio.

## Separability

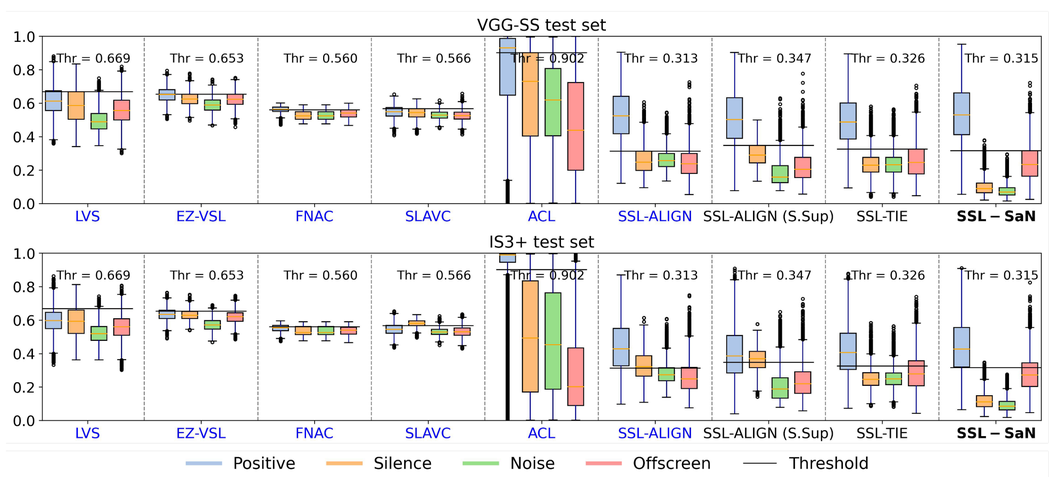
A good VSSL model should **align** positive audio–visual pairs and **separate** negative ones. We introduce a new metric:

$$\text{Sep} = Q_1^+ - Q_3^-$$

where $Q_1^+$ is the 1st quartile of positives and $Q_3^-$ the 3rd quartile of negatives.

Higher values indicate better discrimination between positives and negatives. SSL-SaN achieves the strongest separability across datasets.

## Separability + Universal Threshold




## Ablation Study

We analyze the effect of negative pairs (Silence, Noise) and loss terms ($\mathcal{L}_S$, $\mathcal{L}_N$).

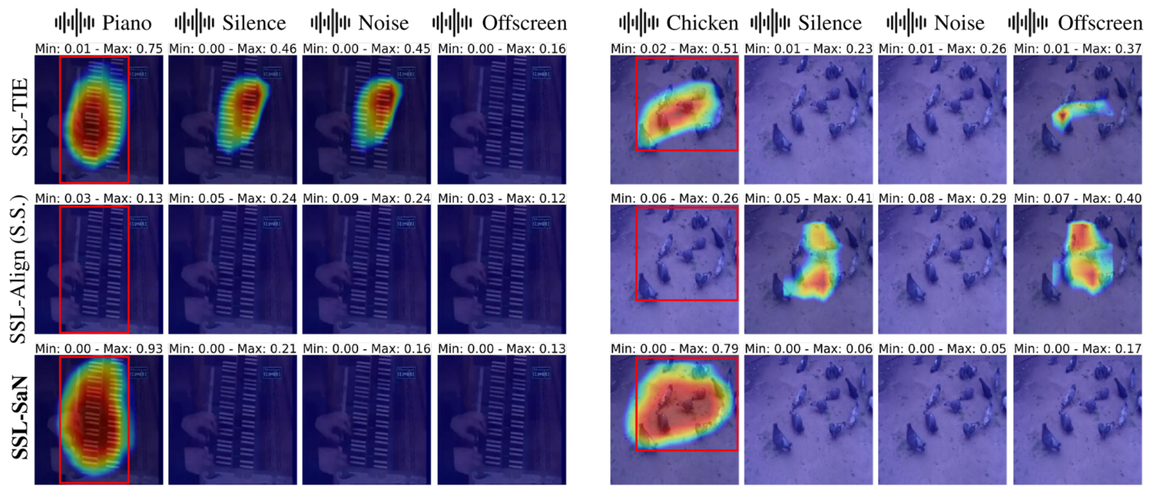
| Test set | S | N | $\mathcal{L}_S$ | $\mathcal{L}_N$ | cIoU$_{Uth}$ ↑ | pIA$_S$ ↓ | pIA$_N$ ↓ | pIA$_O$ ↓ | F$_{LOC}$ ↑ | Sep ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| AVS-Bench S4 | | | | | 31.57 | 2.53 | 2.49 | 1.00 | 47.76 | 0.2384 |
| | ✓ | | | | 31.76 | 2.39 | 1.85 | 0.82 | 48.01 | 0.2350 |
| | ✓ | | ✓ | | 32.49 | 2.54 | 2.58 | 0.81 | 48.80 | 0.2458 |
| | | ✓ | | | 31.97 | 2.26 | 2.59 | 0.88 | 48.22 | 0.2558 |
| | | ✓ | | ✓ | 32.07 | 2.21 | 2.26 | 1.07 | 48.34 | 0.2375 |
| | ✓ | ✓ | | | 32.05 | 1.75 | 1.01 | 0.84 | 48.40 | 0.2393 |
| | ✓ | ✓ | ✓ | ✓ | 32.76 | 0.05 | 0.00 | 0.95 | 49.31 | 0.2479 |

- Best overall metrics occur with **Silence + Noise + ($\mathcal{L}_S + \mathcal{L}_N$)**.
- Offscreen sounds improve with Silence + $\mathcal{L}_S$, but at the cost of positives.
- Separability is highest with Noise (without $\mathcal{L}_N$), second best with full setting.

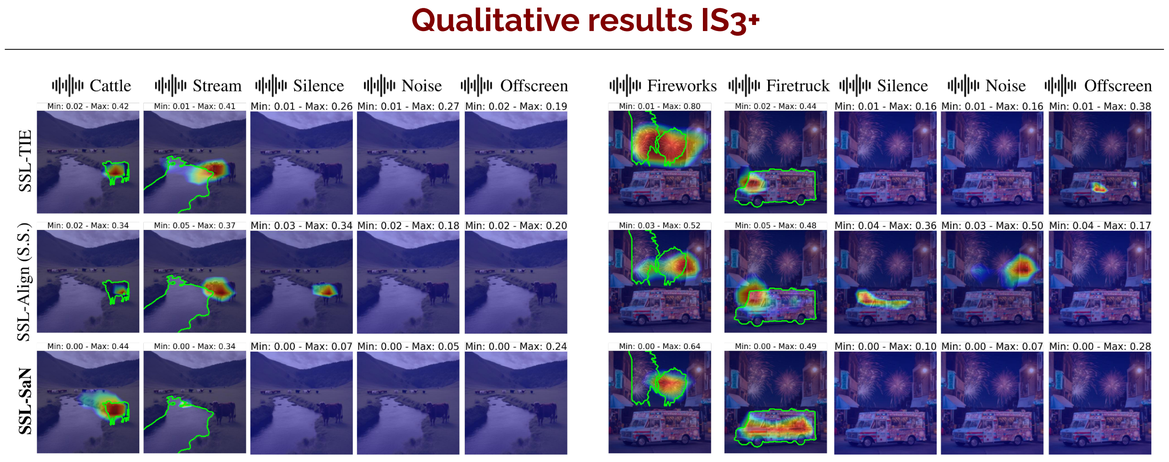
## Quantitative results

| Test set | Model | Self Sup. | Positive audio input cIoU Uth ↑ | Positive audio input cIoU Adap. ↑ | Positive audio input AUC Uth ↑ | Positive audio input AUC Adap. ↑ | Negative audio input Silence pIA ↓ | Silence AUC$_N$ ↑ | Noise pIA ↓ | Noise AUC$_N$ ↑ | Offscreen sound pIA ↓ | Offscreen sound AUC$_N$ ↑ | Global metric F$_{LOC}$ ↑ | Global metric F$_{AUC}$ ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VGG-SS | RCGrad [INTERSPEECH, 2022] | ✘ | 11.71 | 37.04 | 12.08 | 37.26 | 4.42 | 95.80 | 4.42 | 95.80 | 4.41 | 95.82 | 20.86 | 21.45 |
| | LVS [CVPR, 2021] | ✘ | 3.90 | 39.43 | 5.91 | 41.27 | 1.59 | 98.30 | 0.05 | 99.93 | 0.70 | 99.23 | 7.51 | 11.15 |
| | EZ-VSL [ECCV, 2022] | ✘ | 10.41 | 43.85 | 11.97 | 42.86 | 1.83 | 98.09 | 0.37 | 99.60 | 1.80 | 98.08 | 18.84 | 21.34 |
| | FNAC [CVPR, 2023] | ✘ | 17.45 | 47.14 | 18.69 | 44.27 | 4.09 | 95.88 | 4.09 | 95.88 | 3.58 | 96.38 | 29.53 | 31.28 |
| | SLAVC [NEURIPS, 2022] | ✘ | 9.60 | 49.62 | 11.40 | 45.55 | 5.53 | 94.43 | 0.59 | 99.39 | 1.54 | 98.45 | 17.48 | 20.40 |
| | SSL-Align [ICCV, 2023] | ✘ | 34.46 | 56.86 | 34.78 | 49.25 | 2.34 | 97.57 | 1.11 | 98.79 | 1.97 | 97.95 | 51.02 | 51.36 |
| | ACL [WACV 2025] | ✘ | 14.31 | 39.92 | 15.99 | 40.30 | 0.29 | 99.60 | 0.40 | 99.47 | 0.91 | 99.01 | 25.02 | 27.55 |
| | SSL-TIE [ACMMM, 2022] | ✓ | 27.78 | 51.88 | 28.23 | 48.00 | 0.78 | 99.16 | 0.68 | 99.26 | 2.50 | 97.39 | 43.36 | 43.90 |
| | SSL-Align [ICCV, 2023] | ✓ | 25.40 | 53.92 | 25.96 | 47.84 | 1.18 | 98.72 | 0.28 | 99.70 | 0.63 | 99.33 | 40.45 | 41.16 |
| | Ours → SSL-SaN | ✓ | 29.61 | 52.74 | 29.97 | 48.66 | 0.01 | 99.99 | 0.00 | 100.00 | 1.72 | 98.17 | 45.63 | 46.05 |
| IS3+ | RCGrad [INTERSPEECH, 2022] | ✘ | 0.10 | 1.07 | 2.54 | 3.57 | 4.50 | 52.49 | 4.49 | 52.49 | 4.48 | 52.49 | 0.20 | 4.95 |
| | LVS [CVPR, 2021] | ✘ | 2.05 | 14.35 | 4.31 | 26.72 | 1.08 | 98.80 | 0.03 | 99.95 | 0.29 | 99.67 | 4.01 | 8.26 |
| | EZ-VSL [ECCV, 2022] | ✘ | 3.91 | 17.95 | 5.95 | 28.70 | 1.16 | 98.75 | 0.02 | 99.98 | 0.95 | 98.95 | 7.52 | 11.23 |
| | FNAC [CVPR, 2023] | ✘ | 7.62 | 16.84 | 9.37 | 28.53 | 3.43 | 96.54 | 3.43 | 96.54 | 3.22 | 96.73 | 14.13 | 17.08 |
| | SLAVC [NEURIPS, 2022] | ✘ | 6.43 | 18.23 | 8.31 | 28.33 | 18.57 | 81.38 | 0.49 | 99.49 | 3.21 | 96.76 | 12.02 | 15.25 |
| | SSL-Align [ICCV, 2023] | ✘ | 25.63 | 33.36 | 26.29 | 38.11 | 4.33 | 95.49 | 1.17 | 98.67 | 2.30 | 97.58 | 40.58 | 41.39 |
| | ACL [WACV 2025] | ✘ | 44.75 | 63.93 | 45.79 | 64.16 | 0.02 | 99.86 | 1.37 | 98.79 | 0.39 | 99.64 | 61.72 | 62.70 |
| | SSL-TIE [ACMMM, 2022] | ✓ | 16.37 | 17.93 | 17.40 | 28.88 | 0.49 | 99.42 | 0.40 | 99.53 | 3.03 | 96.86 | 28.09 | 29.58 |
| | SSL-Align [ICCV, 2023] | ✓ | 17.11 | 28.90 | 18.30 | 35.99 | 3.33 | 96.45 | 0.55 | 99.42 | 0.70 | 99.23 | 29.15 | 30.86 |
| | Ours → SSL-SaN | ✓ | 19.13 | 18.90 | 19.94 | 30.28 | 0.00 | 100.00 | 0.00 | 100.00 | 2.09 | 97.73 | 32.08 | 33.21 |
| AVS-Bench S4 | RCGrad [INTERSPEECH, 2022] | ✘ | 15.76 | 33.08 | 16.15 | 33.25 | 4.34 | 95.92 | 4.34 | 95.91 | 4.33 | 95.92 | 27.07 | 27.65 |
| | LVS [CVPR, 2021] | ✘ | 6.82 | 21.30 | 8.54 | 30.68 | 2.16 | 97.73 | 0.07 | 99.91 | 0.65 | 99.28 | 12.76 | 15.72 |
| | EZ-VSL [ECCV, 2022] | ✘ | 11.14 | 19.57 | 12.56 | 30.81 | 0.71 | 99.24 | 0.52 | 99.45 | 1.40 | 98.53 | 20.04 | 22.29 |
| | FNAC [CVPR, 2023] | ✘ | 18.57 | 22.85 | 19.55 | 33.14 | 5.79 | 94.17 | 5.79 | 94.17 | 3.17 | 96.81 | 31.07 | 32.42 |
| | SLAVC [NEURIPS, 2022] | ✘ | 9.07 | 22.58 | 10.80 | 33.13 | 3.72 | 96.24 | 1.17 | 98.81 | 1.42 | 98.55 | 16.60 | 19.45 |
| | SSL-Align [ICCV, 2023] | ✘ | 33.04 | 30.73 | 33.09 | 39.32 | 4.81 | 95.13 | 1.04 | 98.88 | 1.48 | 98.45 | 49.36 | 49.41 |
| | ACL [WACV 2025] | ✘ | 51.34 | 65.75 | 49.57 | 64.11 | 0.00 | 99.99 | 0.03 | 99.98 | 0.32 | 99.72 | 67.82 | 66.26 |
| | SSL-TIE [ACMMM, 2022] | ✓ | 28.40 | 31.24 | 28.64 | 38.60 | 2.90 | 97.08 | 2.69 | 97.25 | 1.37 | 98.56 | 44.00 | 44.29 |
| | SSL-Align Senocak et al. [2018] [ICCV, 2023] | ✓ | 31.75 | 32.30 | 31.94 | 39.35 | 0.87 | 99.03 | 0.16 | 99.81 | 0.48 | 99.49 | 48.14 | 48.35 |
| | Ours → SSL-SaN | ✓ | 32.76 | 33.87 | 32.86 | 41.11 | 0.05 | 99.95 | 0.00 | 100.00 | 0.95 | 98.96 | 49.31 | 49.42 |

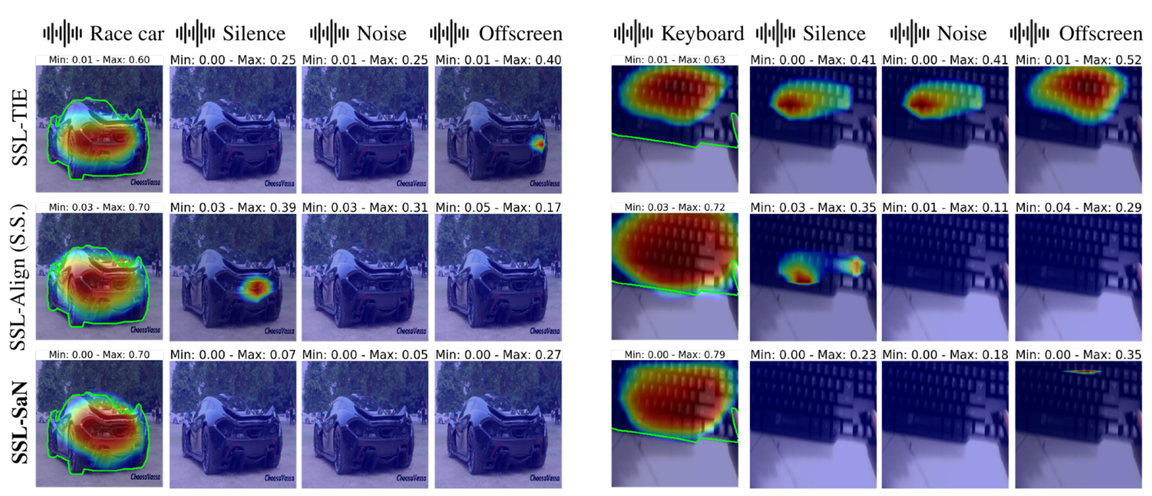
## Qualitative results VGG-SS




## Qualitative results IS3+




## Qualitative results AVS-Bench S4




## References

Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16867–16876, 2021.

Xavier Juanola, Gloria Haro, and Magdalena Fuentes. A critical assessment of visual sound source localization models including negative audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.

Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3742–3753, 2022.

Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems*, 35:37524–37536, 2022a.

Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, pages 218–234. Springer, 2022b.

Sooyoung Park, Arda Senocak, and Joon Son Chung. Can clip help sound source localization? In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5711–5720, 2024.

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Aligning sight and sound: Advanced sound source localization through audio-visual alignment. *arXiv preprint arXiv:2407.13676*, 2024.

Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2023.

Ho-Hsiang Wu, Magdalena Fuentes, Prem Seetharaman, and Juan Pablo Bello. How to listen? rethinking visual sound localization. *arXiv preprint arXiv:2204.05156*, 2022.

## Project Page

xavijuanola.github.io/SSL-SaN

EXCELENCIA MARÍA DE MAEZTU

Universitat Pompeu Fabra Barcelona — Departament d'Enginyeria

GOBIERNO DE ESPAÑA — MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES



xavier.juanola@upf.edu