TFM topic: **Robust visual sound source localization and cross-modal retrieval**

Team: Gloria Haro and Xavier Juanola (Universitat Pompeu Fabra)
Magdalena Fuentes (New York University)

Visual Sound Source Localization focuses on finding where a sound comes from in an image by combining audio and visual information. While current models perform well in simple scenarios, they lack robustness when faced with overlapping sources, ambient noise, or off-screen acoustic events [1]. This TFM aims to improve existing approaches by using self-supervised and contrastive learning techniques within a multimodal framework, enabling the model to learn from natural audio-visual relationships without relying on detailed annotations. The goal is to achieve more robust and reliable sound localization in realistic scenarios with mixed and noisy sound sources. Furthermore, the learned audio-visual representation will be applied to cross-modal retrieval.

The objective of this TFM is to extend the work presented in [2]. The proposed method will incorporate more expressive encoders and will leverage the Tri-map formulation introduced in [3] (Figure 1), which reframes sound source localization using positive, negative, and ambiguous regions to explicitly handle uncertainty in audio-visual correspondences. This TFM aims to extend this idea by applying an analogous Tri-map strategy to the audio domain, exploiting its temporal and frequential dimensions (Figure 2). In Figure 2, the red regions highlight the time-frequency bins associated with two distinct sound sources within the STFT (Short-Time Fourier Transform) domain. By modeling these confident regions in time–frequency representations, we expect the approach to provide more reliable learning signals for self-supervised training, further improving robustness in complex scenarios with overlapping sounds.

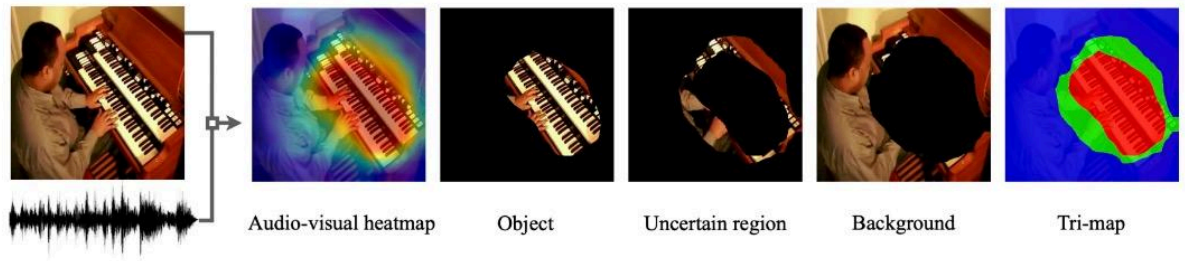

Figure 1

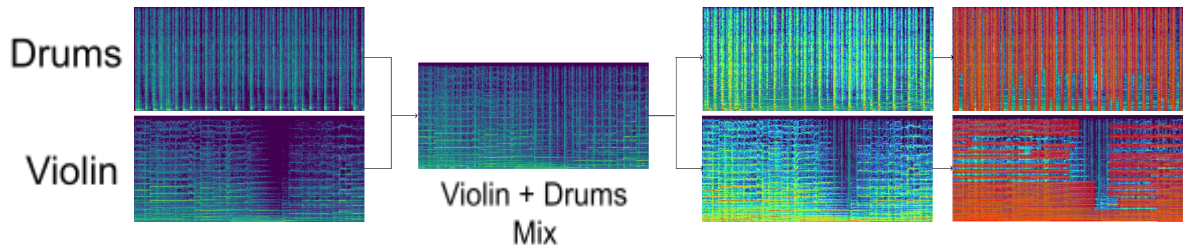

Figure 2

References:

[1] X. Juanola, G. Haro, M. Fuentes. A Critical Assessment of Visual Sound Source Localization Models Including Negative Audio, ICASSP 2025. https://xavijuanola.github.io/vssleval/

[2] X. Juanola, G. Morais, M. Fuentes, G. Haro.Learning from Silence and Noise for Visual Sound Source Localization, BMVC 2025. https://xavijuanola.github.io/SSL-SaN/

[3] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, A. Zisserman. Localizing Visual Sounds the Hard Way. CVPR 2021. https://www.robots.ox.ac.uk/~vgg/research/lvs/