A Brief Analysis of SLAVC method for Sound Localization

Xavier Juanola Molet
Gloria Haro

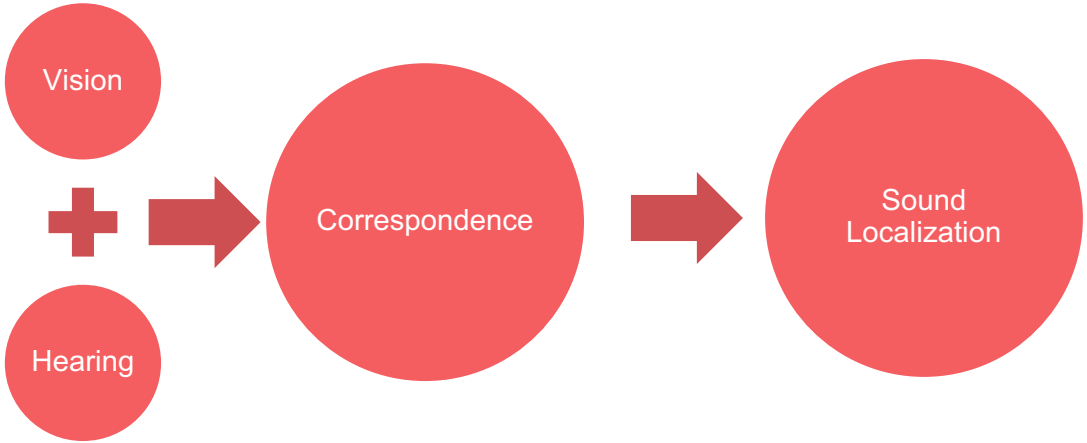
IPOL Journal · Image Processing On Line

Universitat Pompeu Fabra Barcelona

Index

Index

1. INTRODUCTION



Fundamental aspect of human perception

- Navigate our environment
- Communicate
- Respond to potential threats

1. INTRODUCTION

Hershey and Movellan [1], Fisher et al. [2] and Kidron et al. [3] → Model based correspondences between visual features and audio features

Arandjelovic and Zisserman [4] Senocak et al. [5] and Hu et al. [6] → Used Contrastive Learning to localize objects aligning visual and audio representations

1. INTRODUCTION

Hershey and Movellan [1], Fisher et al. [2] and Kidron et al. [3] → Model based correspondences between visual features and audio features

Arandjelovic and Zisserman [4] Senocak et al. [5] and Hu et al. [6] → Used Contrastive Learning to localize objects aligning visual and audio representations

---

2 Major flaws: →

1. Rely on Early-stopping to avoid overfitting.
2. Assumed all sound sources are present in the scene

→ Incorrectly identifying false positives

→ Recent Works identify silent objects and off-screen sounds (Hu et al. [7] and Liu et al. [8]

1. INTRODUCTION

Hershey and Movellan [1], Fisher et al. [2] and Kidron et al. [3] → Model based correspondences between visual features and audio features

Arandjelovic and Zisserman [4] Senocak et al. [5] and Hu et al. [6] → Used Contrastive Learning to localize objects aligning visual and audio representations

---

2 Major flaws: →

1. Rely on Early-stopping to avoid overfitting.
2. Assumed all sound sources are present in the scene

→ Incorrectly identifying false positives

→ Recent Works identify silent objects and off-screen sounds (Hu et al. [7] and Liu et al. [8]

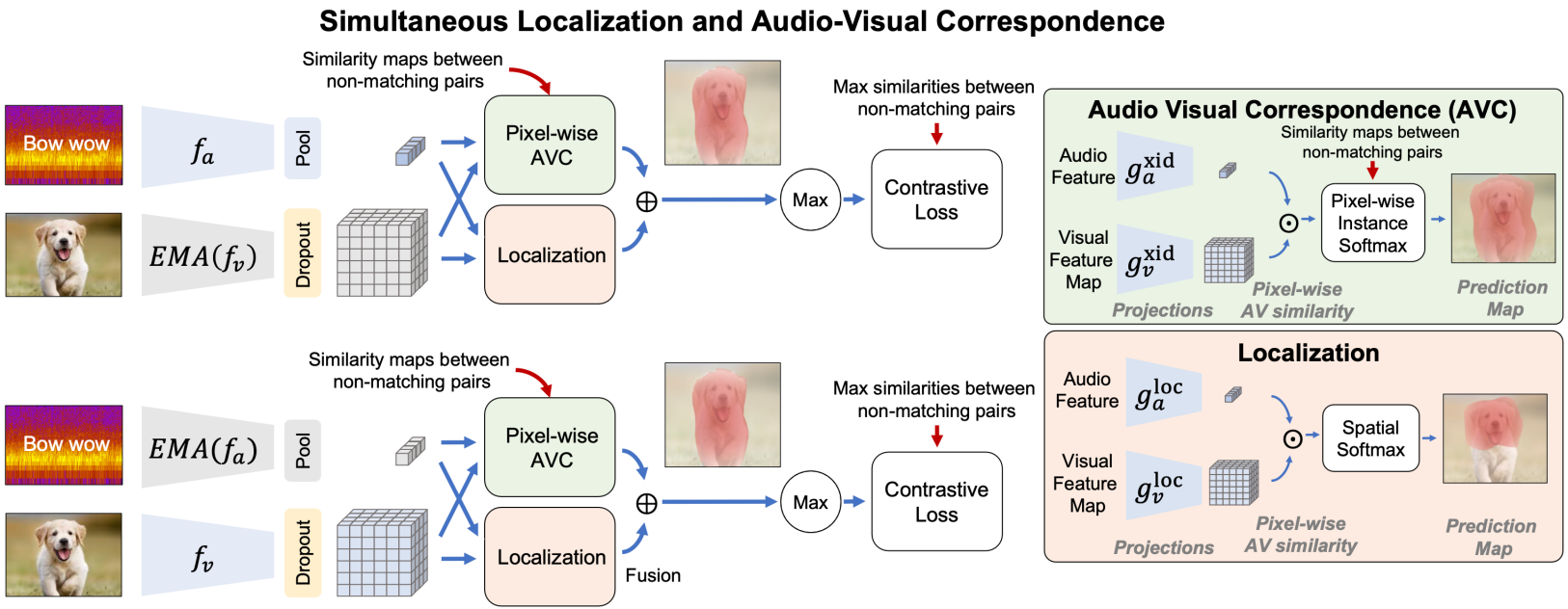
SLAVC —solves→

- Overfitting → Heavy visual dropout + Momentum encoders
- Hallucinating visible sound objects →
  1. Visual Sound Localization Term
  2. Audio - Visual Correspondence Term

Index

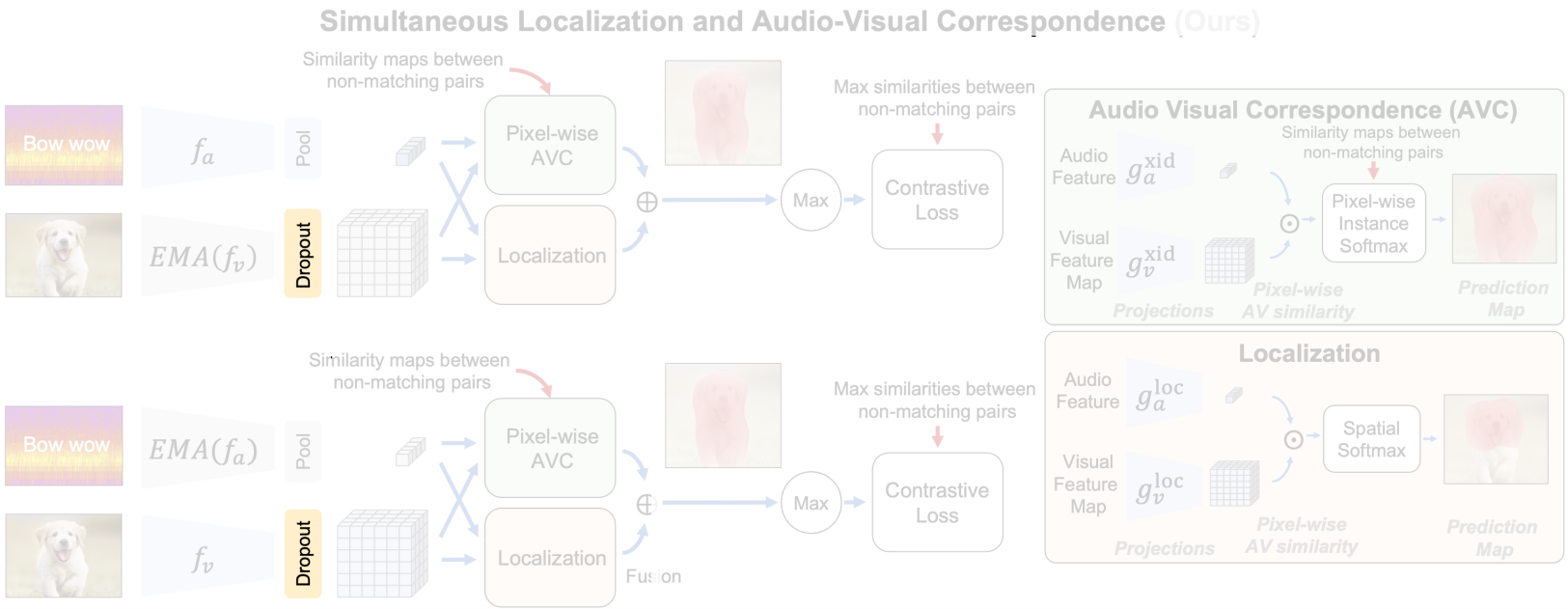
2. METHOD

SLAVC



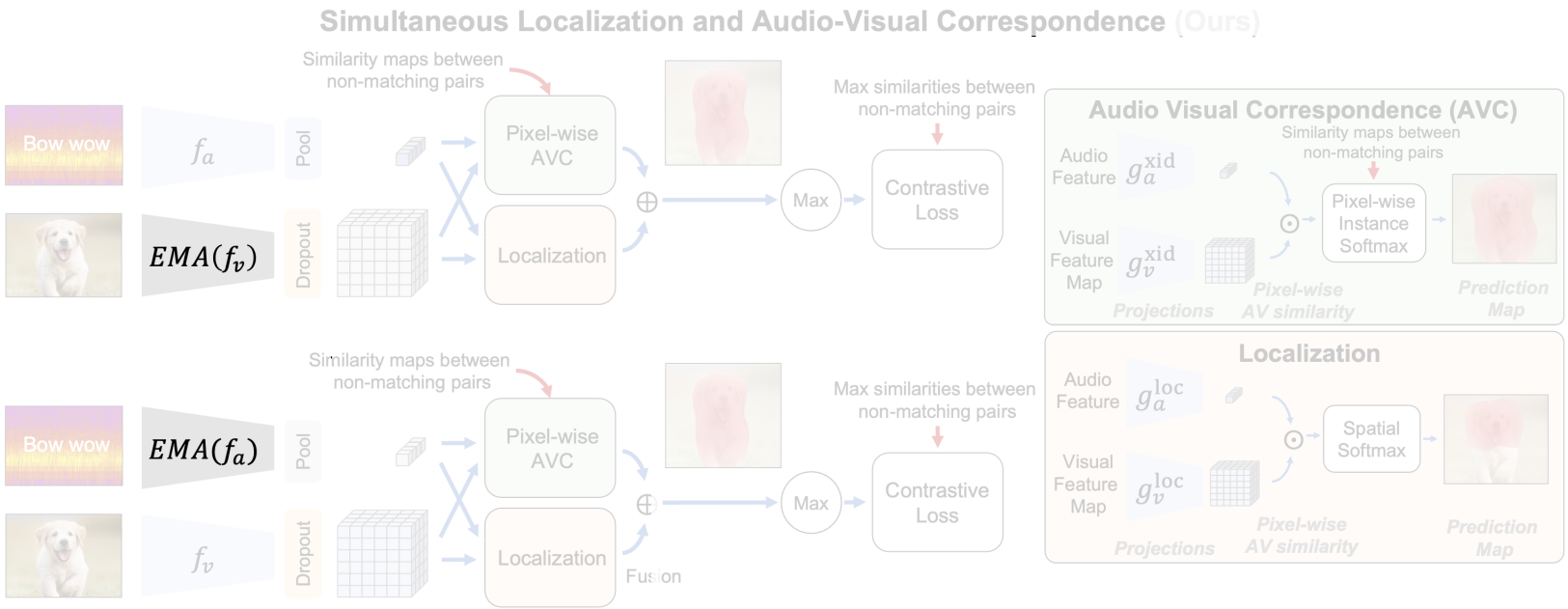
2. METHOD

SLAVC → Overfitting → Heavy visual **dropout** + Momentum encoders



2. METHOD

SLAVC → Overfitting → Heavy visual dropout + **Momentum encoders**

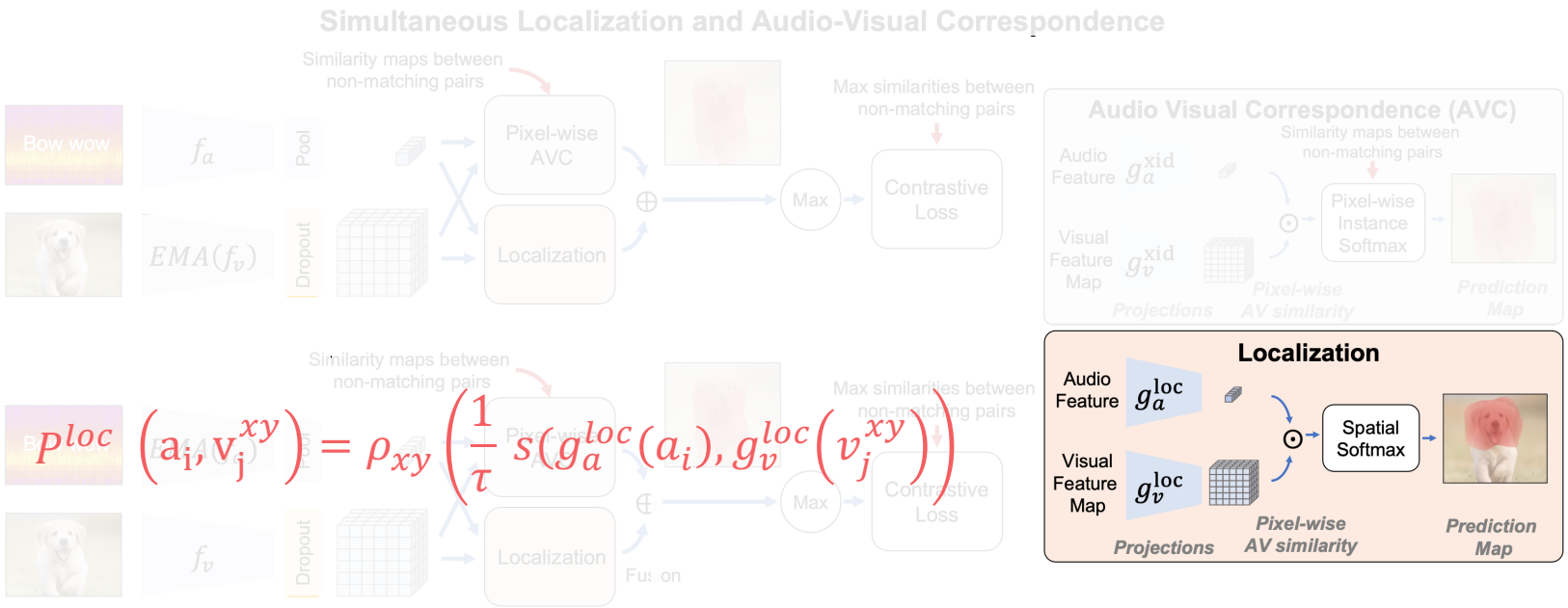


2. METHOD

SLAVC →

Overfitting → Heavy visual dropout + Momentum encoders

Hallucinating visible sound objects → 1. **Visual Sound Localization Term**



$$P^{loc}\left(a_i, v_j^{xy}\right) = \rho_{xy}\left(\frac{1}{\tau} s\left(g_a^{loc}(a_i), g_v^{loc}\left(v_j^{xy}\right)\right)\right)$$

2. METHOD

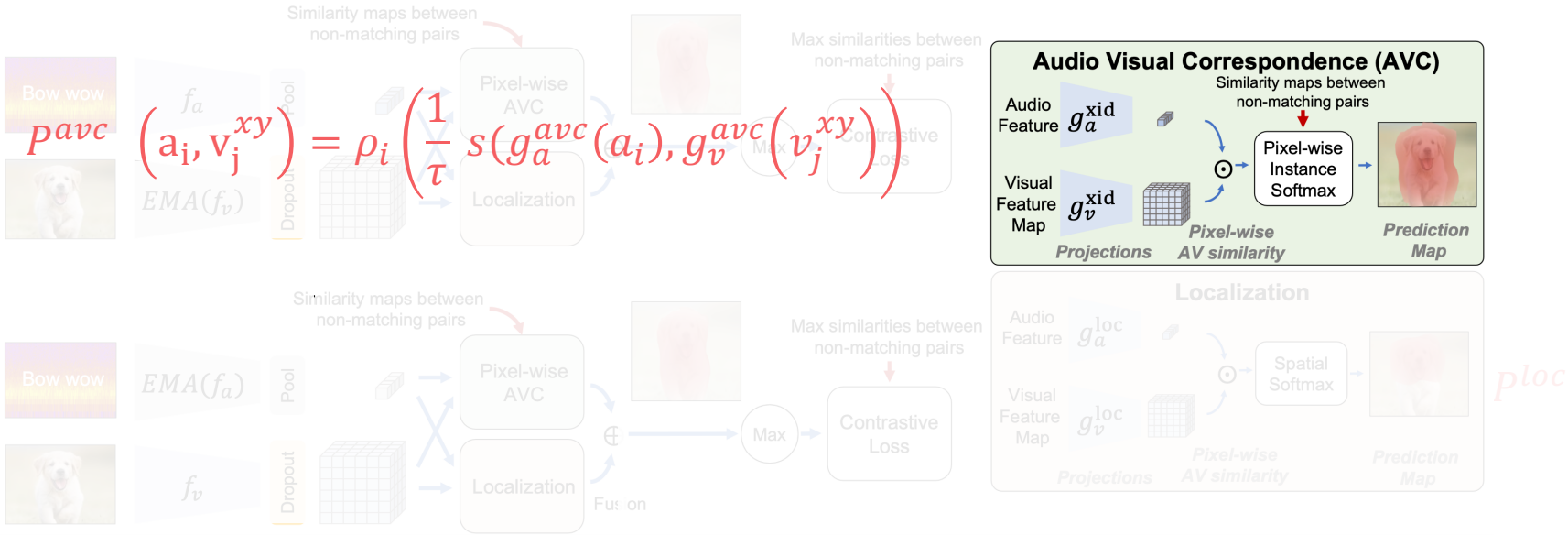
SLAVC →

Overfitting → Heavy visual dropout + Momentum encoders

Hallucinating visible sound objects →

1. Visual Sound Localization Term
2. **Audio - Visual Correspondence Term**

Simultaneous Localization and Audio-Visual Correspondence



$$p^{avc}\left(a_i, v_j^{xy}\right) = \rho_i\left(\frac{1}{\tau} s\left(g_a^{avc}(a_i), g_v^{avc}\left(v_j^{xy}\right)\right)\right)$$

2. METHOD

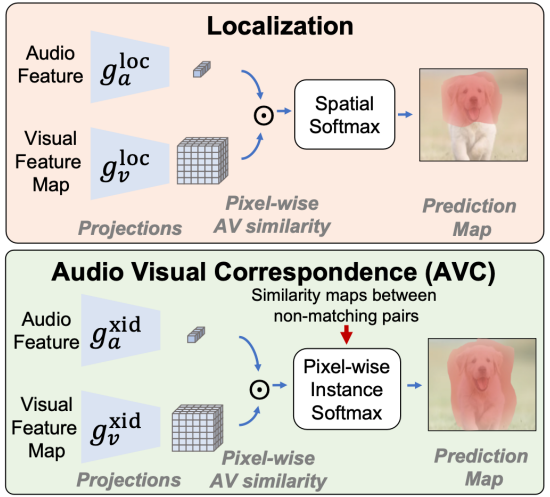
SLAVC

Overfitting → Heavy visual dropout + Momentum encoders

Hallucinating visible sound objects →
1. Visual Sound Localization Term
2. Audio - Visual Correspondence Term

$$P^{loc}\left(a_i, v_j^{xy}\right) = \boldsymbol{\rho_{xy}}\left(\frac{1}{\tau}\, s(g_a^{loc}(a_i), g_v^{loc}\left(v_j^{xy}\right)\right)$$

$$P^{avc}\left(a_i, v_j^{xy}\right) = \boldsymbol{\rho_i}\left(\frac{1}{\tau}\, s(g_a^{avc}(a_i), g_v^{avc}\left(v_j^{xy}\right)\right)$$



2. METHOD

SLAVC

Overfitting → Heavy visual dropout + Momentum encoders

Hallucinating visible sound objects →
1. Visual Sound Localization Term
2. Audio - Visual Correspondence Term

$$P^{SLAVC}\left(a_i, v_j^{xy}\right) = P^{loc}\left(a_i, v_j^{xy}\right) \cdot P^{avc}\left(a_i, v_j^{xy}\right)$$

2. METHOD

SLAVC

Overfitting → Heavy visual dropout + Momentum encoders

Hallucinating visible sound objects →
1. Visual Sound Localization Term
2. Audio - Visual Correspondence Term

$$P^{SLAVC}\left(a_i, v_j^{xy}\right) = P^{loc}\left(a_i, v_j^{xy}\right) \cdot P^{avc}\left(a_i, v_j^{xy}\right)$$

$$S_{VSL}^{xy} = \alpha S_{SLAVC}^{xy} + (1-\alpha) S_{OGL}^{xy}$$

S_{OGL}^{xy} → Object Guided Localization (OGL)

S_{VSL}^{xy} → Visual Sound Localization (VSL)

2. METHOD

SLAVC

Overfitting → Heavy visual dropout + Momentum encoders

Hallucinating visible sound objects →
1. Visual Sound Localization Term
2. Audio - Visual Correspondence Term

$$P^{SLAVC}\left(a_i, v_j^{xy}\right) = P^{loc}\left(a_i, v_j^{xy}\right) \cdot P^{avc}\left(a_i, v_j^{xy}\right)$$

$$S_{VSL}^{xy} = \alpha S_{SLAVC}^{xy} + (1-\alpha) S_{OGL}^{xy}$$

Input **SLAVC** **OGL** **VSL** $(\alpha = 0.5)$

Index

3. DEMO

IPOL Journal · Image Processing On Line

HOME · ABOUT · ARTICLES · PREPRINTS · WORKSHOPS · NEWS · SEARCH

A Closer Look at Weakly-Supervised Audio-Visual Source Localization demo

Article | Demo | Archive

Please cite the reference article if you publish results obtained with this online demo.

Description

Demo of the paper 'A Closer Look at Weakly-Supervised Audio-Visual Source Localization'.

Select input(s)

Upload data

Input(s)

Parameters

Restablecer

alpha 0,5 Max: 1 Min: 0

Level of importance of SLAVC with respect of OGL in VSL. VSL(x,y) = alpha * SLAVC(x,y) + (1 - alpha) * OGL(x,y)

Run

feeds & twitter • sitemap • contact • privacy policy • ISSN: 2105-1232 • DOI: 10.5201/ipol
IPOL and its contributors acknowledge support from September 2010 to August 2015 by the European Research Council (advanced grant Twelve Labours n°246961).
IPOL is also supported by ONR grant N00014-14-1-0023, CNES (MISS project), FUI 18 Plein Phare project, and ANR-DGA project ANR-12-ASTR-0035.
IPOL is maintained by Centre Borelli, ENS Paris-Saclay, DMI, Universitat de les Illes Balears, and Fing, Universidad de la República.

3. DEMO

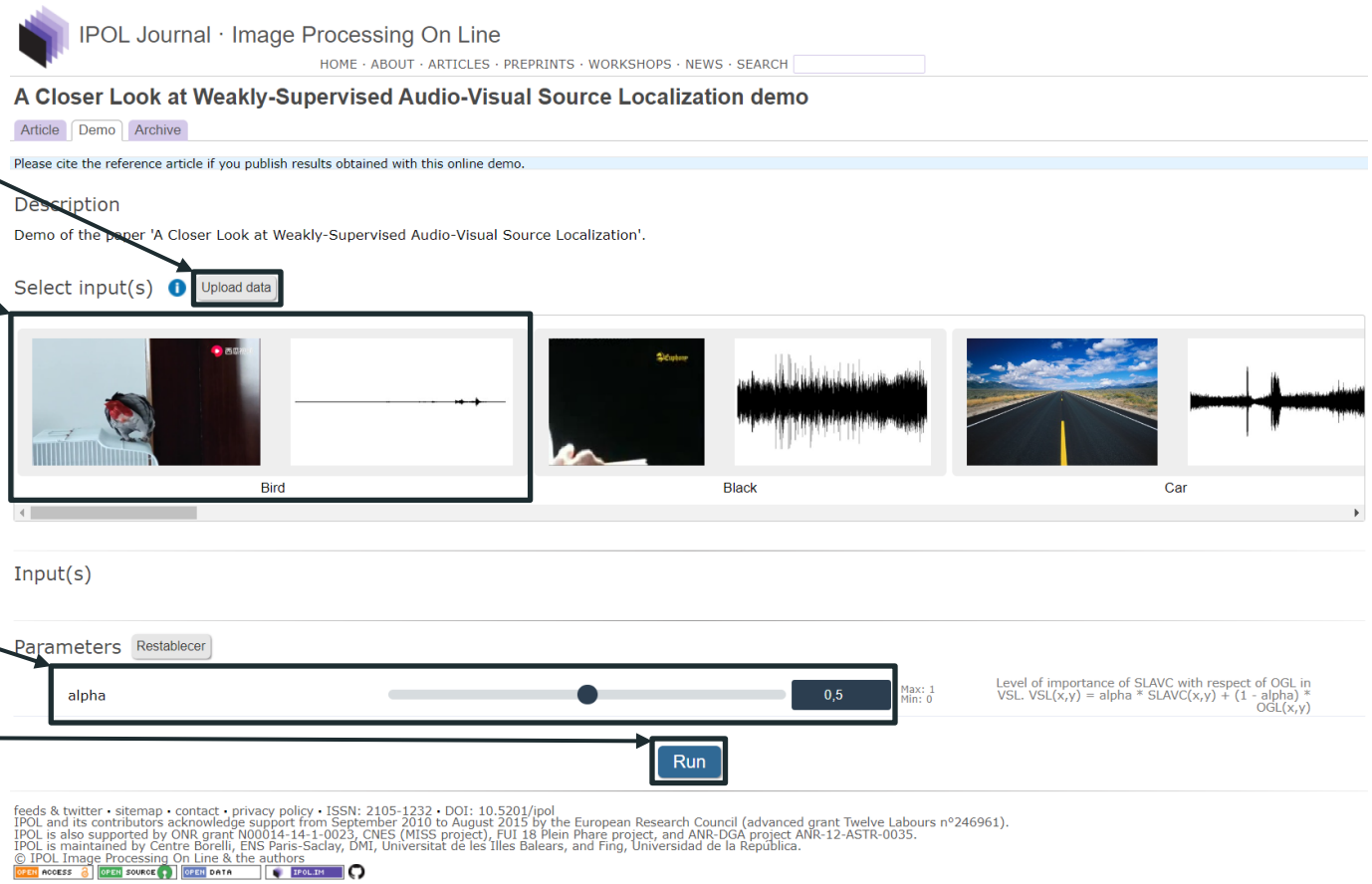
Upload image and audio file

Select pair of image audio from the ones provided

Select α value:

$$S_{VSL}^{xy} = \alpha \cdot S_{SLAVC}^{xy} + (1-\alpha) \cdot S_{OGL}^{xy}$$

Run the demo



3. DEMO

Results

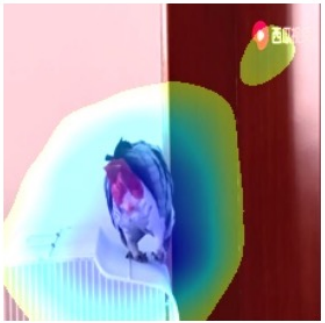
Input

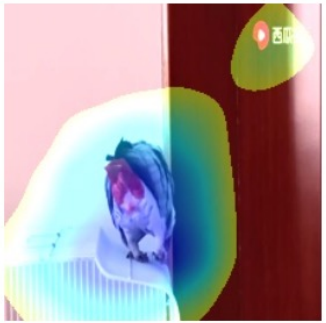
Simultaneous Localization and Audio-Visual Correspondence (SLAVC)

Object Guided Localization (OGL)

Visual Sound Localization (VSL)

☑ Compare





Input

Simultaneous Localization and Audio-Visual Correspondence (SLAVC)

Object Guided Localization (OGL)

Visual Sound Localization (VSL)

Zoom 1x

Colorbar

Zoom 1x

Index

4. EXPERIMENTS

1. Impact of α
2. Easy cases
3. Difficulties on Visual Sound Localization
   - 3.1 Mixture of sounds
   - 3.2 Small objects
   - 3.3 Silent objects
   - 3.4 Off-screen sounds
   - 3.5 Different objects of the same type

4. EXPERIMENTS

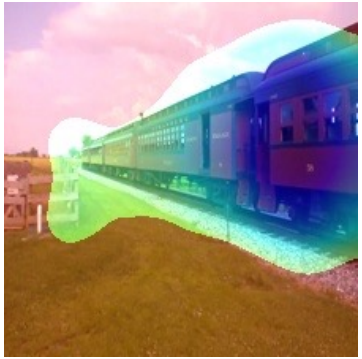
Impact of α

$$S_{VSL}^{xy} = \alpha S_{SLAVC}^{xy} + (1-\alpha) S_{OGL}^{xy}$$

Input

SLAVC

OGL



$\alpha = 0.10$

$\alpha = 0.25$

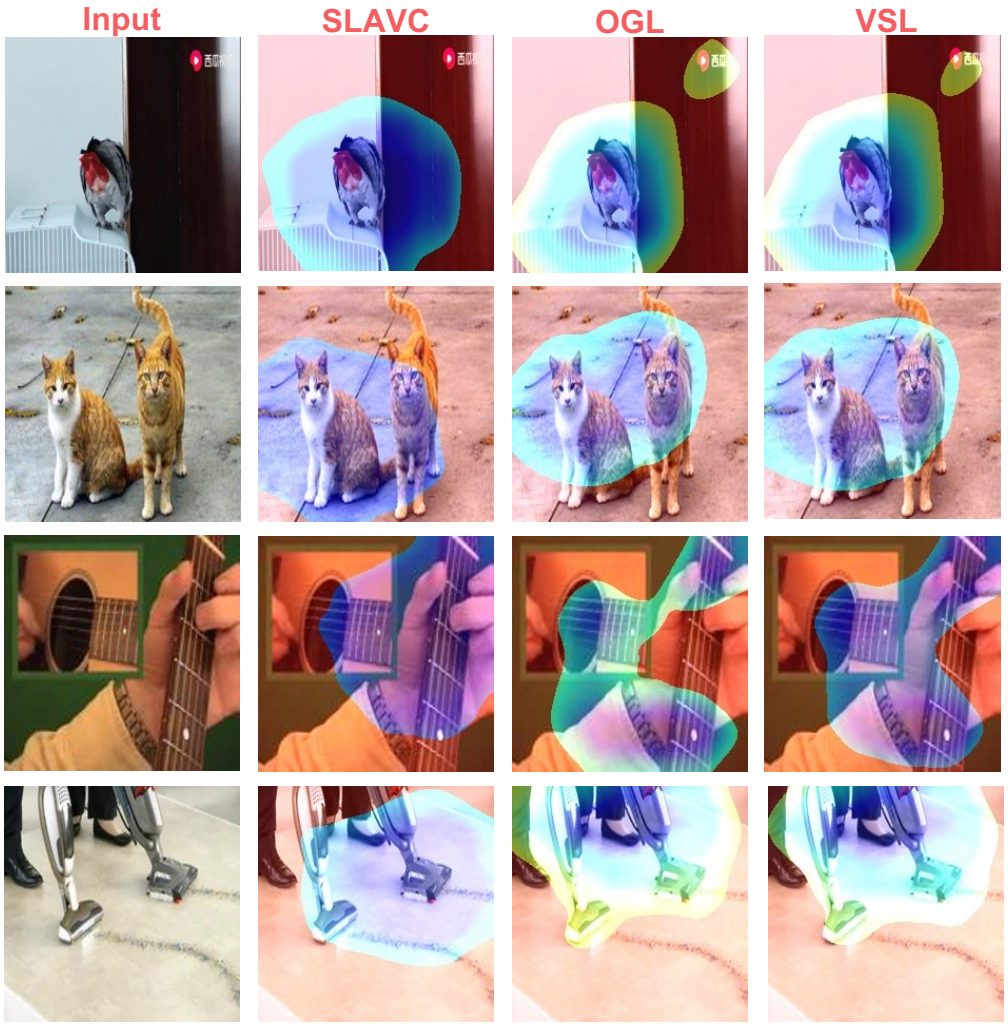
$\alpha = 0.50$

$\alpha = 0.75$

$\alpha = 0.90$

4. EXPERIMENTS

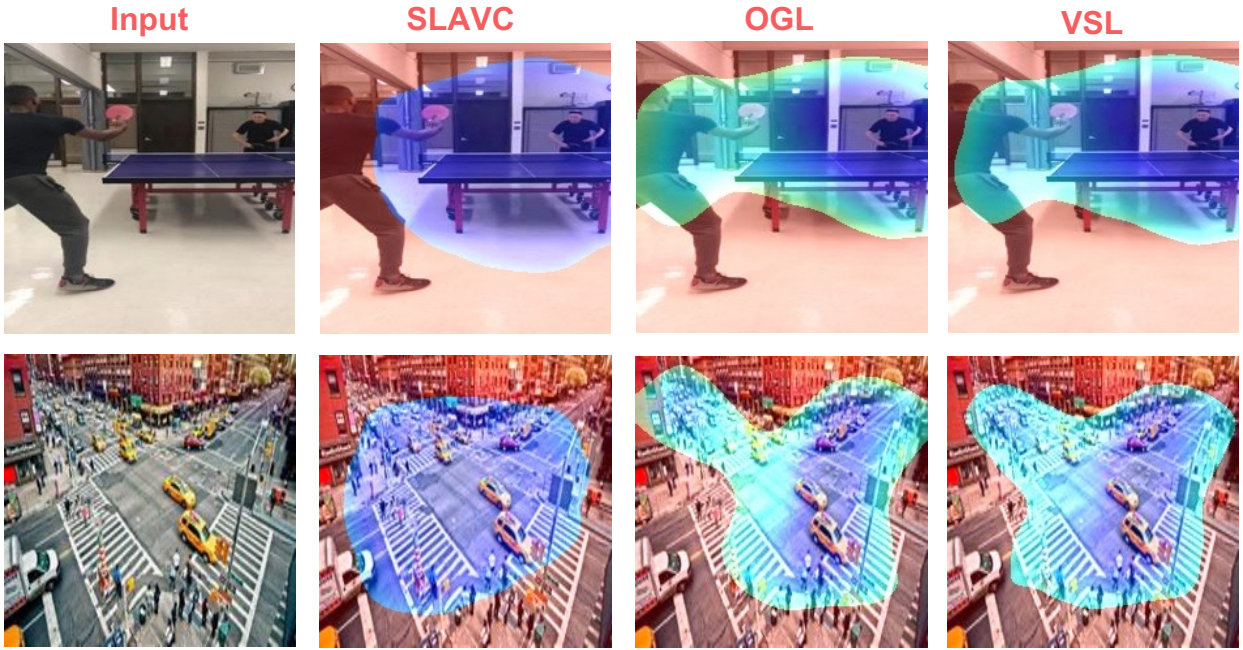
Easy cases



4. EXPERIMENTS

Difficult cases:

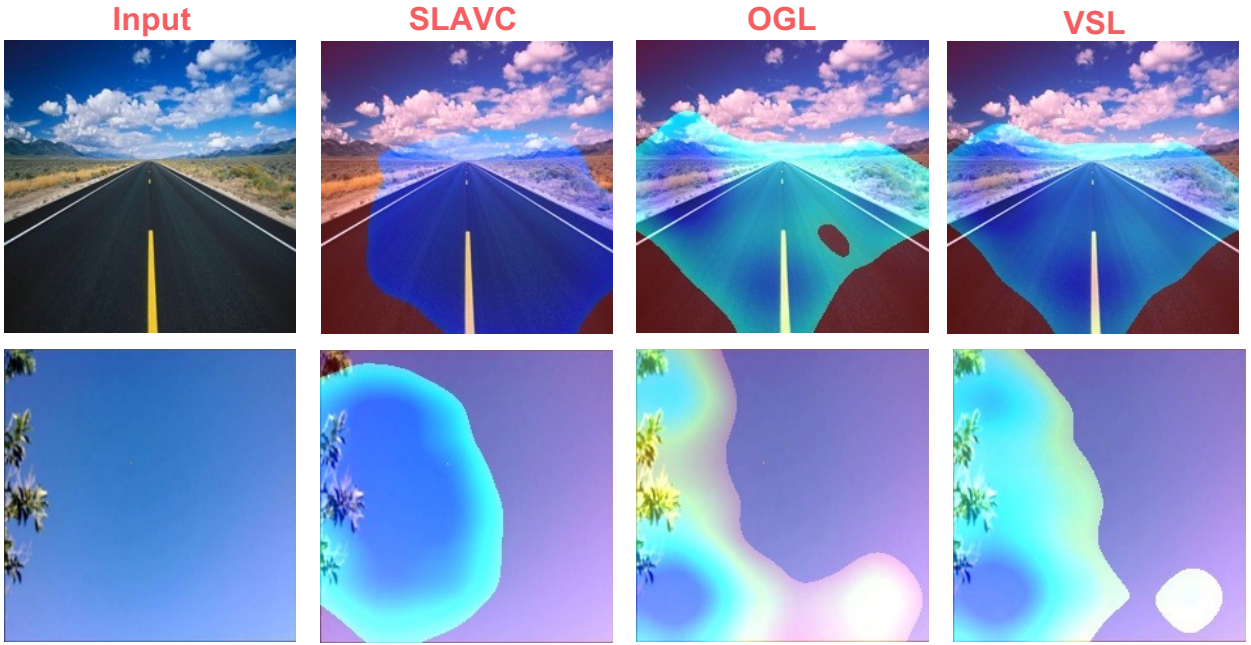
1. Mixture of sounds



4. EXPERIMENTS

Difficult cases:

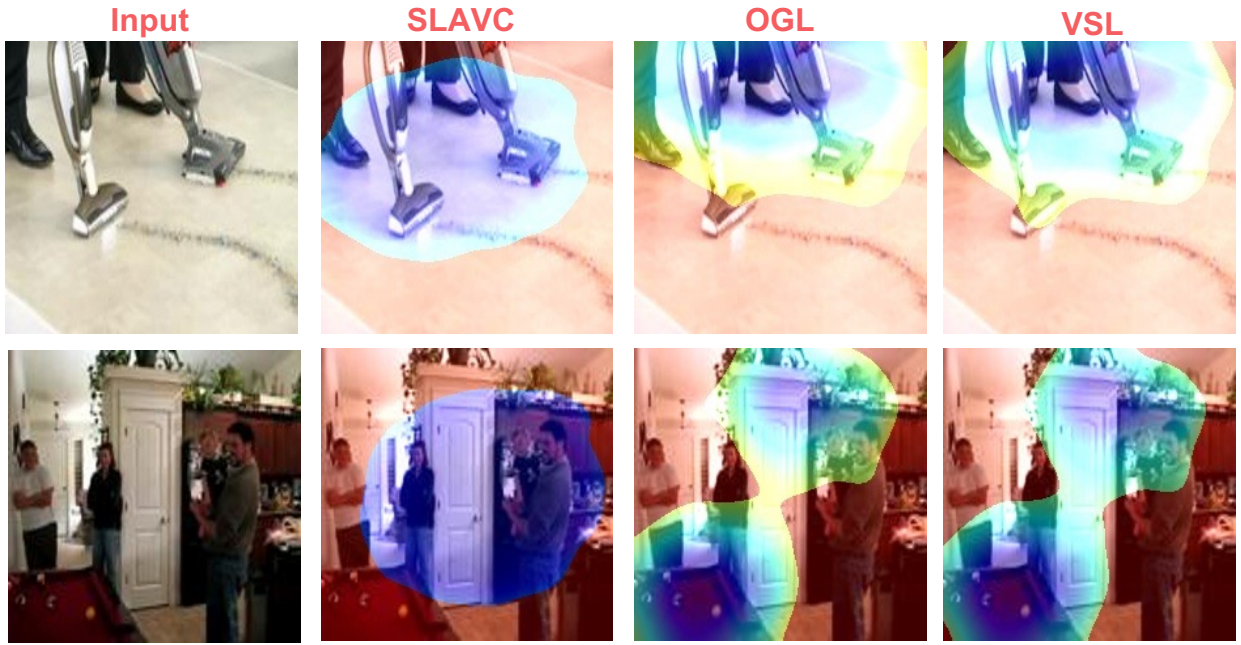
2. Small objects



4. EXPERIMENTS

Difficult cases:

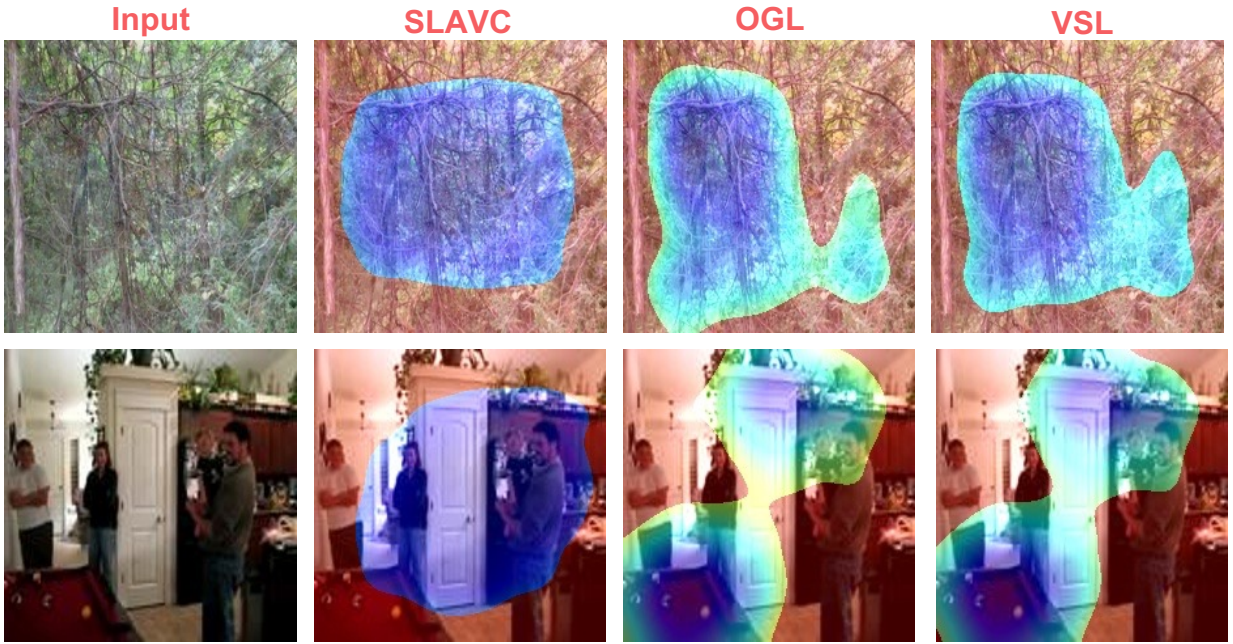
3. Silent objects



4. EXPERIMENTS

Difficult cases:

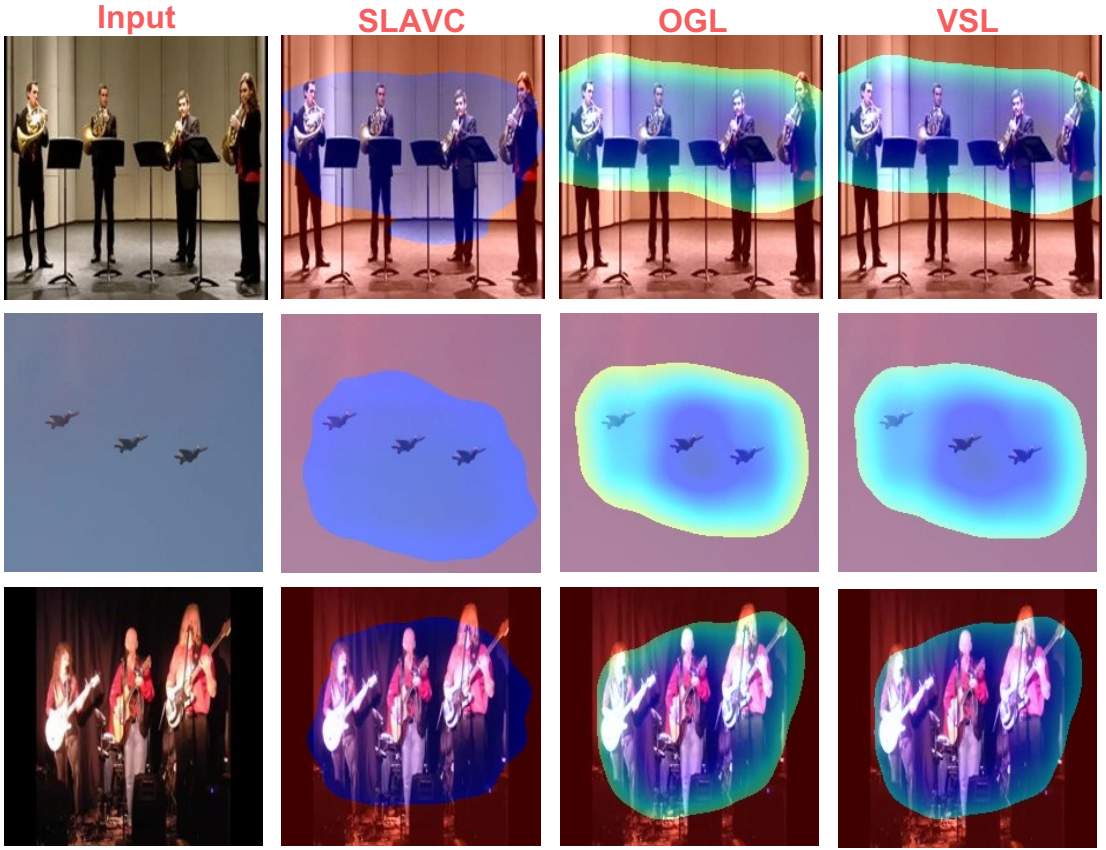
4. Off-screen sounds



4. EXPERIMENTS

Difficult cases:

5. Different objects of the same type



Index

5. PROPOSED IMPROVEMENTS

- ✓ Good results on many cases
- ✗ Mixture of sounds
- ✗ Small objects
- ✗ Silent objects
- ✗ Off-screen sounds
- ✗ Different objects of the same type

1. Image → Videos → Learn motion cues
2. Audio and Visual prototypes used to define proper filters to be applied in the localization map (Liu et al. [8]

Bibliography

1. John Hershey and Javier Movellan, Audio vision: Using audio-visual synchrony to locate sounds, in Advances in Neural Information Processing Systems, S. Solla, T. Leen, and K. M¨uller, eds., vol. 12, MIT Press, 1999.
2. John W Fisher III, Trevor Darrell, William Freeman, and Paul Viola, Learning joint statistical models for audio-visual fusion and segregation, Advances in neural information processing systems, 13 (2000)
3. Einat Kidron, Yoav Y Schechner, and Michael Elad, Pixels that sound, in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, IEEE. 2005, pp. 88–95.
4. Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In ICCV
5. 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, 2018, pp. 4358–4366.
6. Di Hu, Feiping Nie, and Xuelong Li, Deep multimodal clustering for unsupervised audiovisual learning, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9248–9257.
7. Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou, Discriminative sounding objects localization via self-supervised audiovisual matching, Advances in Neural Information Processing Systems, 33 (2020), pp. 10077–10087.
8. Xian Liu, Rui Qian, Hang Zhou, Di Hu, Weiyao Lin, Ziwei Liu, Bolei Zhou, and Xiaowei Zhou, Visual sound localization in the wild by cross-modal interference erasing, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 1801–1809.

Questions?