Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Quan Nguyen,[1] **Julius Richter**,[2] Mikko Lauri,[1] Timo Gerkmann,[2] Simone Frintrop[1]

Improving Mix-and-Separate Training in Audio-Visual Sound Source Separation with an Object Prior

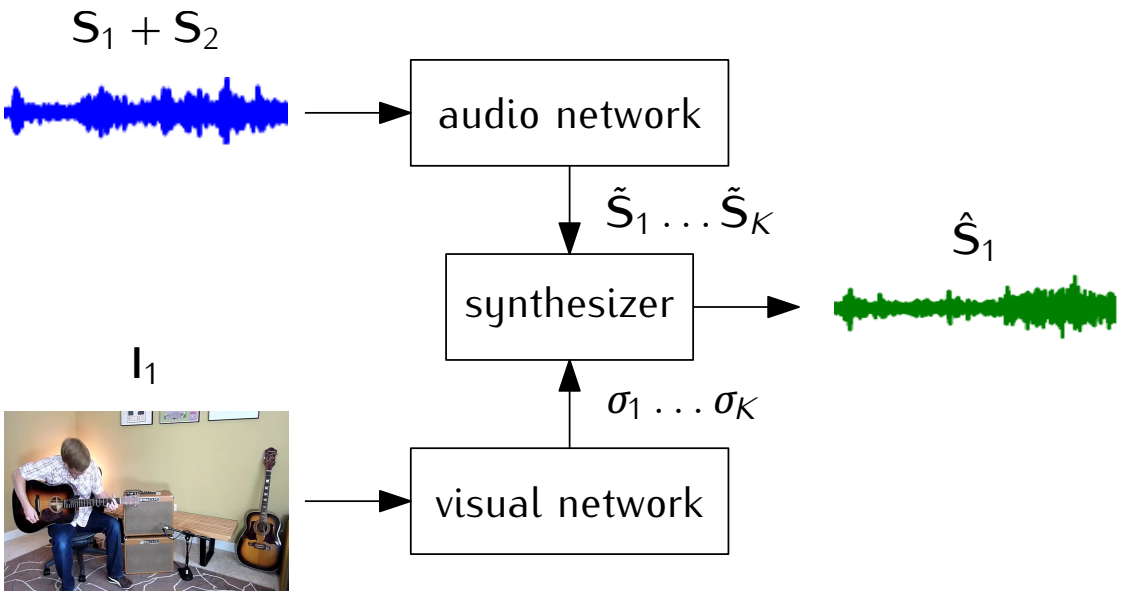
ICPR 2020

Universität Hamburg
[1]Computer Vision Group
[2]Signal Processing Group

January 10–15, 2020

Introduction

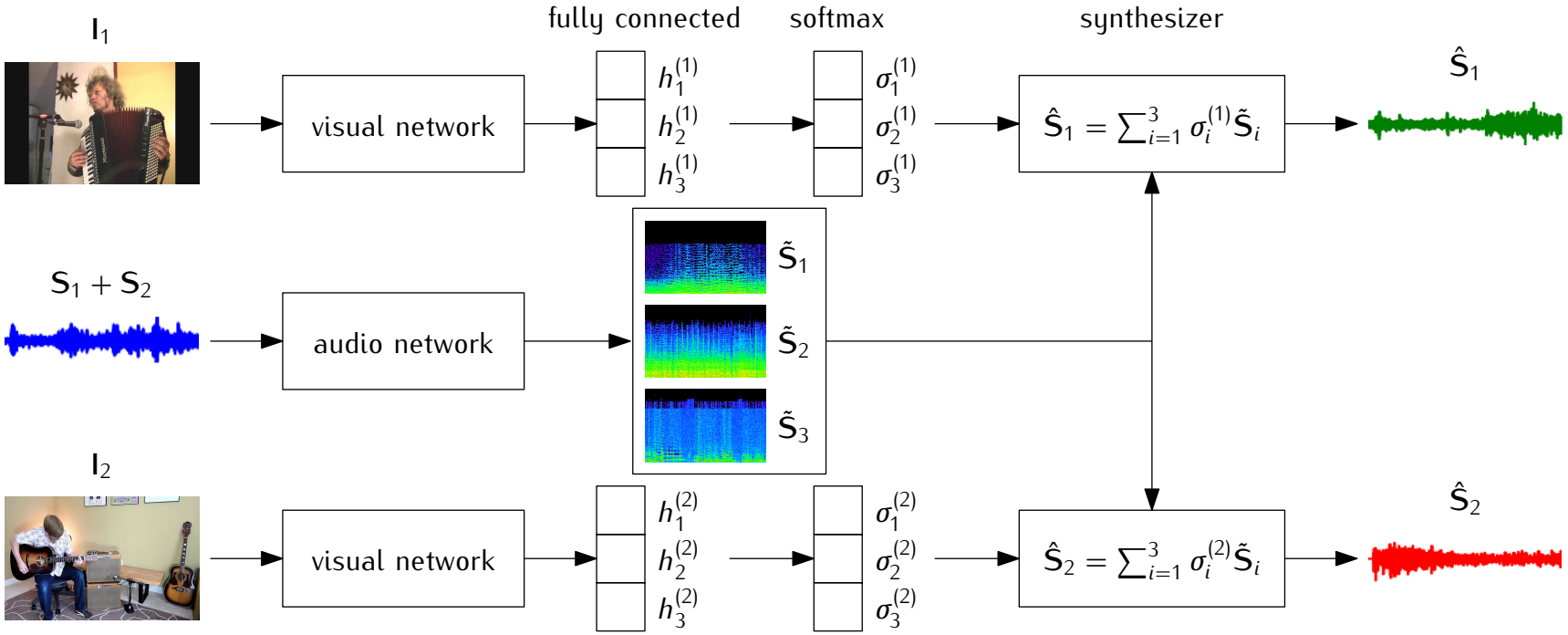
- **Task:** Audio-Visual Sound Source Separation



$$\hat{\mathbf{S}}_1 = \sum_{i=1}^{K} \sigma_i \tilde{\mathbf{S}}_i$$

Background

- Mix-and-separate training paradigm (e.g. in *PixelPlayer*[1])



[1] H. Zhao et al. "The Sound of Pixels". In: *The European Conference on Computer Vision (ECCV)*. Sept. 2018.

Motivation

PixelPlayer[1]:

- Self-supervised method achievses good performance
- Does not automatically learn the 1-to-1 correspondence between the audio and visual channels

Our proposal:

- Weakly-supervised method called *Object-Prior*
- Late-fusion architecture trained first on object classification and then on sound source separation

[1] H. Zhao et al. "The Sound of Pixels". In: *The European Conference on Computer Vision (ECCV)*. Sept. 2018.

Object-Prior Method

- Each audio channel gets assigned to the sound of one object

Two-step training:

1. Video network is trained to recognize the instrument type with one-hot encoding labels
2. Audio network is trained while video network is frozen

- This approach is weakly-supervised rather than self-supervised since the first step requires human supervision

Results

Metrics: SDR, SIR, SAR: the higher, the better[2]

	SDR	SIR	SAR
PixelPlayer [1]	7.30	11.9	**11.9**
Object-Prior (Ours)	**8.92**	**14.49**	11.44

Table: Performance on the MUSIC test set[1] (75 videos)

	SDR	SIR	SAR
PixelPlayer [1]	1.66	3.58	**11.5**
Object-Prior (Ours)	**6.58**	**12.33**	9.28

Table: Performance on the AudioSet-SingleSource test set[3] (165 videos)

[2] E. Vincent, R. Gribonval, and C. Févotte. "Performance measurement in blind audio source separation". In: *IEEE transactions on audio, speech, and language processing* 14.4 (2006), pp. 1462–1469.

[1] H. Zhao et al. "The Sound of Pixels". In: *The European Conference on Computer Vision (ECCV)*. Sept. 2018.

[3] R. Gao, R. Feris, and K. Grauman. "Learning to Separate Object Sounds by Watching Unlabeled Video". In: *ECCV*. 2018.

Conclusion

- Proposed a weakly-supervised model for audio-visual sound source separation

Experiments showed:

- Incorporating prior information about the object type improves the sound source quality measured in SDR and SIR
- The *Object-Prior* method achieves state-of-the-art performance