End-to-End Sound Source Separation Conditioned on Instrument Labels

Olga Slizovskaia*[1], Leo Kim*[2], Gloria Haro[1], Emilia Gómez[1,3]

[1]Pompeu Fabra University, [2]University of Waterloo, [3]Joint Research Centre (EC)

*equal contribution, corresponding to: olga.slizovskaia@upf.edu

Source Separation for Unknown Number of Sources

Target cases: bands, ensembles, orchestras

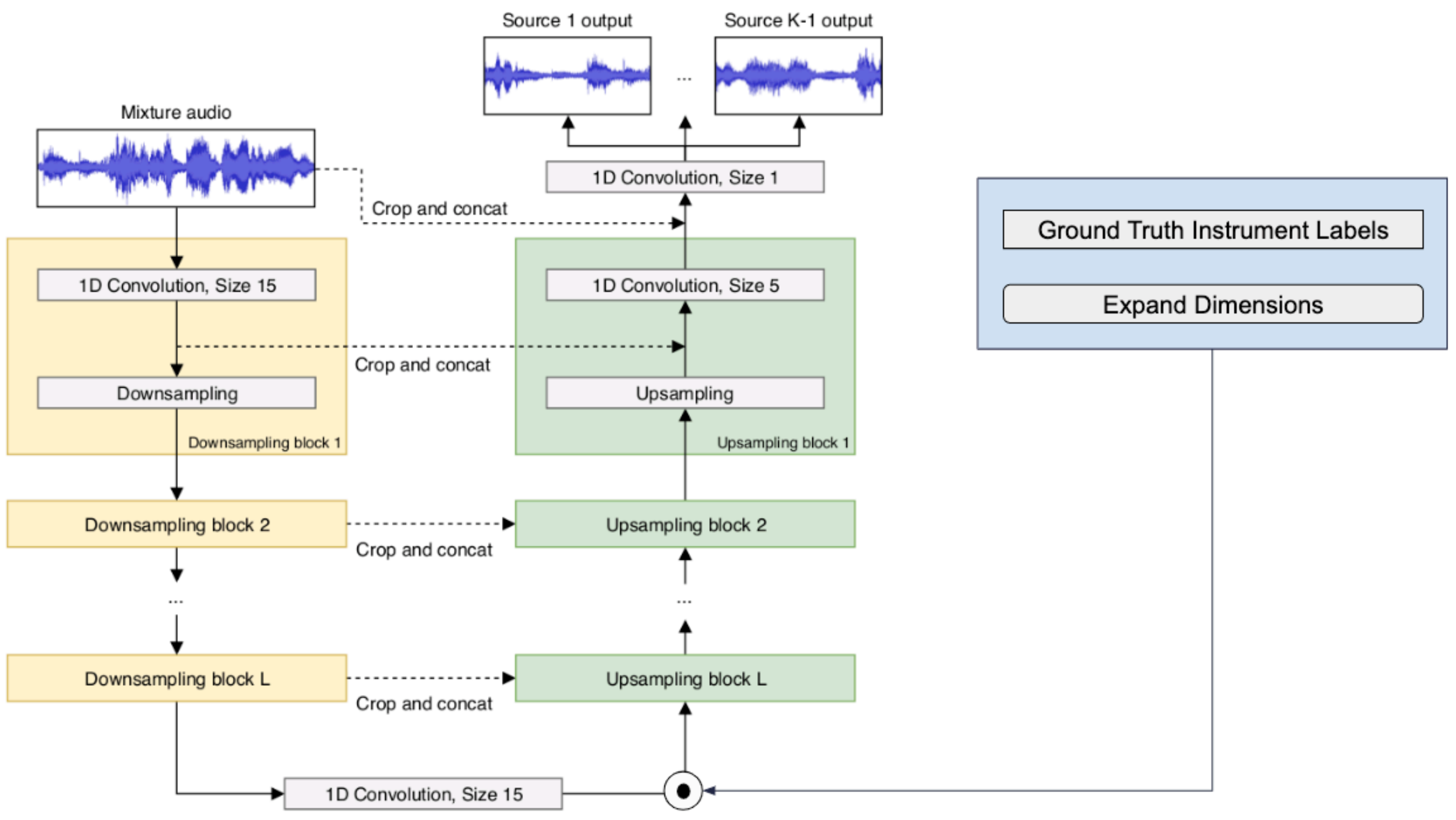
Base architecture: Wave-U-Net [1]

Extension: no predefined number of sources in the mix, multiplicative conditioning with instrument labels

Key features: end-to-end, autoencoder, convolutional, skip connections, upsampling with context

Outlook: extended conditioning for audio-visual and score-informed source separation.

Conditioned Expanded Wave-U-Net Architecture



Mixture audio

Source 1 output

...

Source K-1 output

1D Convolution, Size 1

Crop and concat

1D Convolution, Size 15

1D Convolution, Size 5

Downsampling

Crop and concat

Upsampling

Downsampling block 1

Upsampling block 1

Downsampling block 2

Crop and concat

Upsampling block 2

...

Downsampling block L

Crop and concat

Upsampling block L

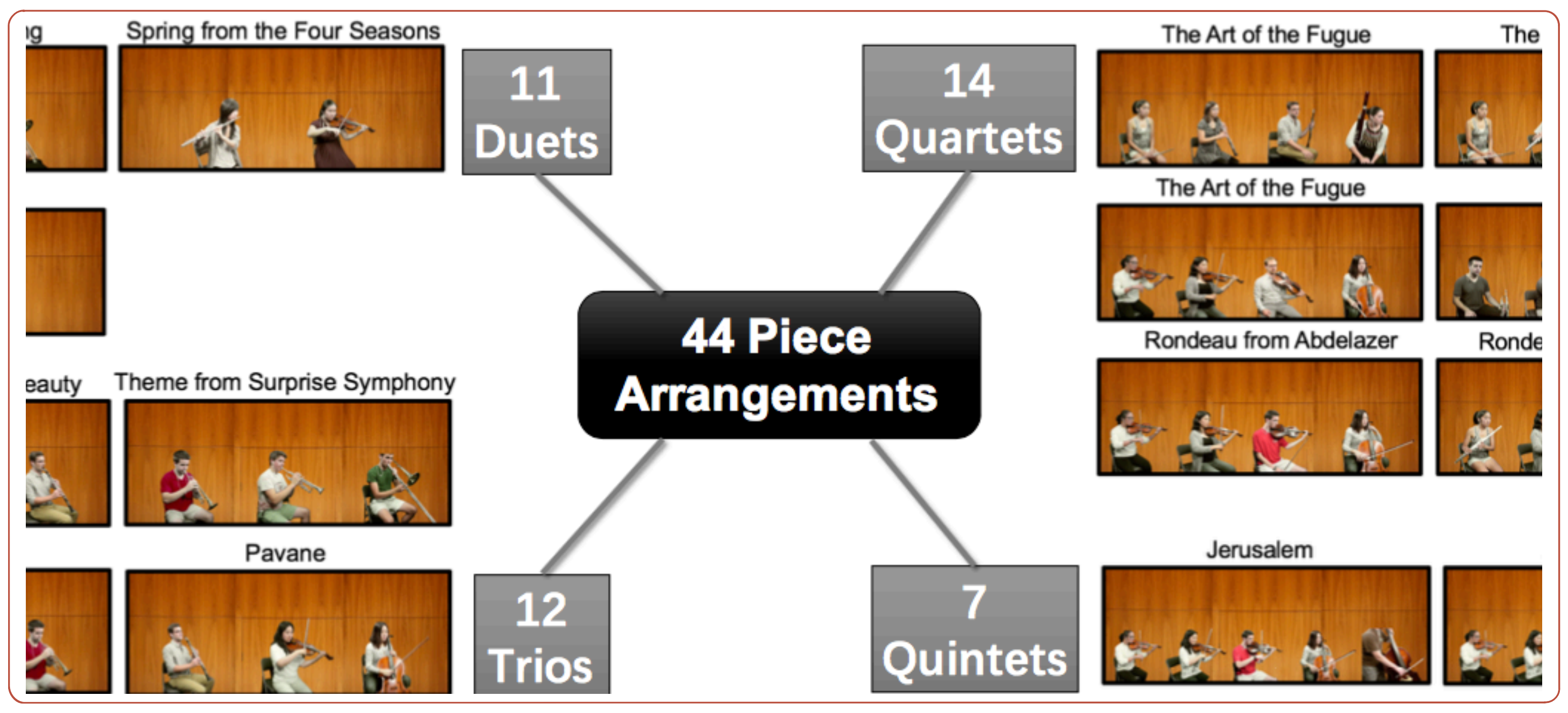
1D Convolution, Size 15

Ground Truth Instrument Labels

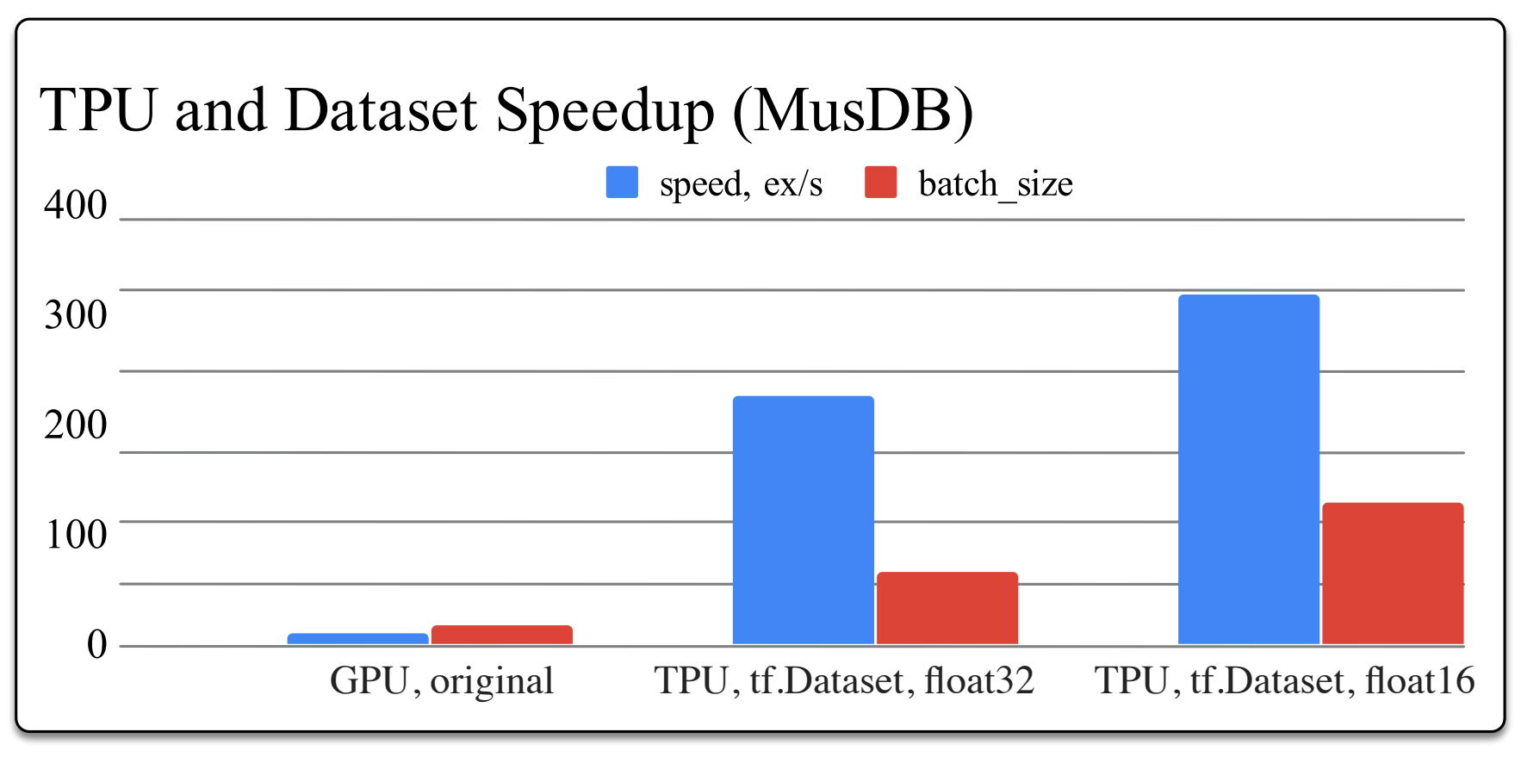
Expand Dimensions

The architecture and the image are adapted from the original Wave-U-Net paper [1].

Multi-Modal URMP Dataset [2]



Faster Training with TPUs

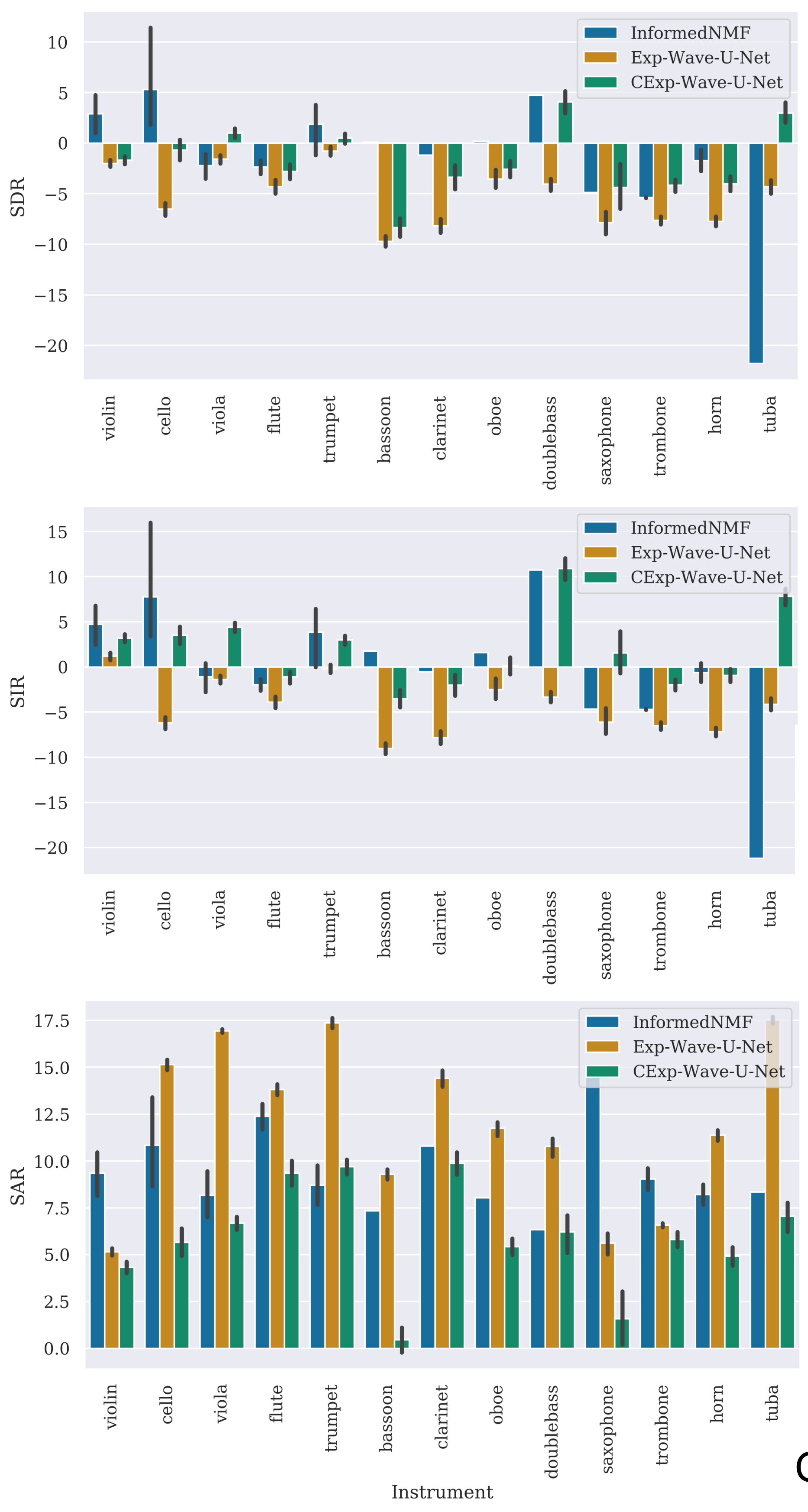


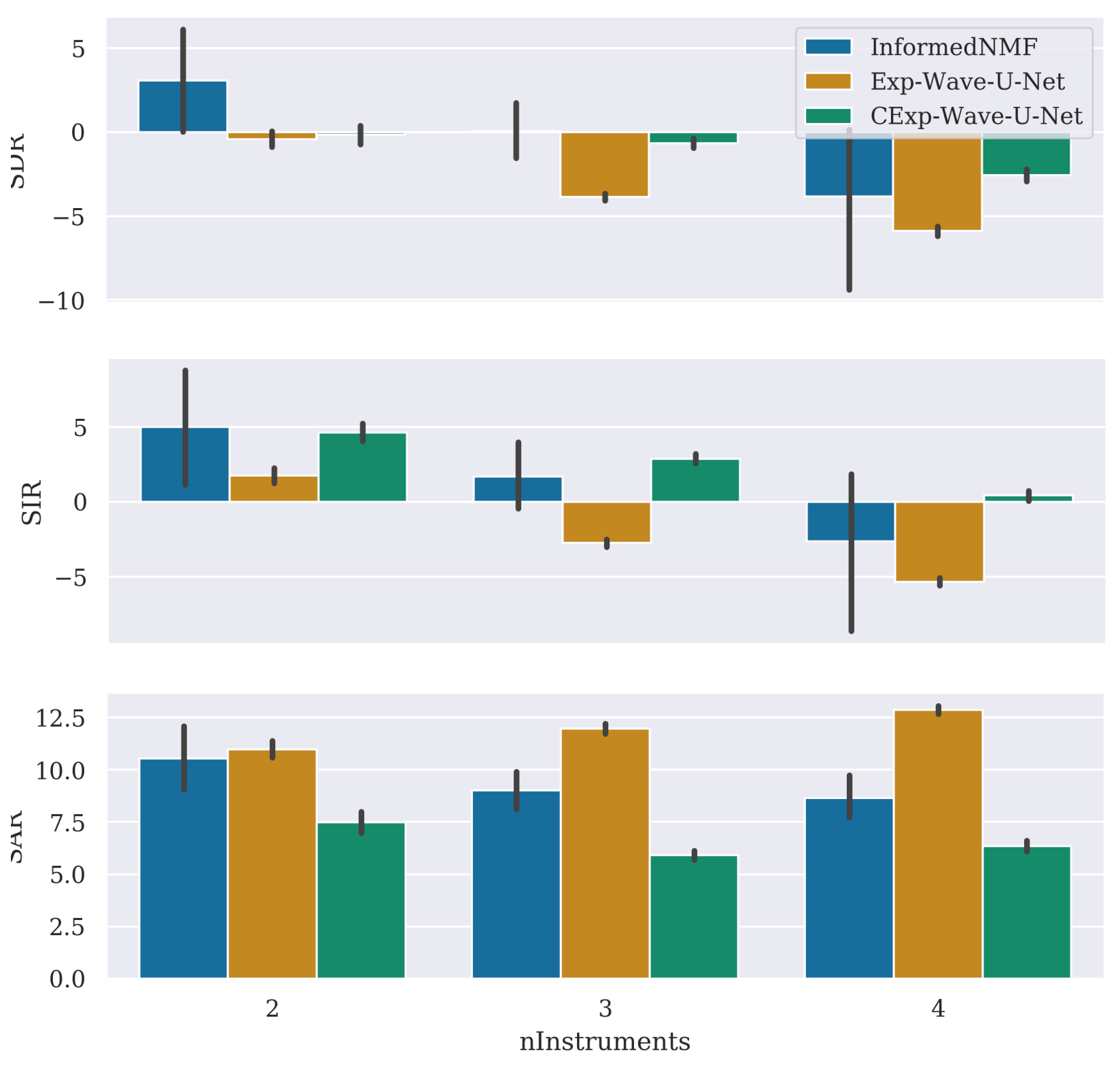
GitHub

veleslavia/vimss

veleslavia/vimss-torch

Results





Method	SDR	SIR	SAR
InformedNMF [3]	**-0.16**	1.42	9.31
Exp-Wave-U-Net	-4.12	-3.06	**12.18**
CExp-Wave-U-Net	-1.37	**2.16**	6.36

Model	nSources	SDR	SIR	SAR
InformedNMF[3]	2	3.08	4.98	10.55
	3	0.07	1.69	9.01
	4	-3.84	-2.62	8.65
Exp-Wave-U-Net	2	-0.42	1.75	10.98
	3	-3.85	-2.74	11.97
	4	-5.90	-5.33	12.87
CExp-Wave-U-Net	2	-0.16	4.62	7.48
	3	-0.68	2.88	5.91
	4	-2.56	0.44	6.35

Qualitative examples https://goo.gl/e18F41

Discussion

- Evaluation is problematic because some sources are silent (we can't estimate with the standard metrics how well the model discards unwanted sources)
- Qualitative examples demonstrate that (C)Exp-Wave-U-Net outputs are more quiet for the undesired sources
- The complexity of the task increases with the number of sources
- CExp-Wave-U-Net performs better in terms of SIR
- CExp-Wave-U-Net performs better than other methods while the number of sources increases
- Exp-Wave-U-Net and CExp-Wave-U-Net are fairly competitive to InformedNMF despite being end-to-end models without explicitly specified timbral models for each instrument

References

[1] D. Stoller, S. Ewert, S. Dixon, et al., "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," 19th International Society for Music Information Retrieval Conference (ISMIR), 2018.

[2] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, vol. PP, 12, 2016.

[3] J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, and F. J. Rodríguez-Serrano, "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, pp. 184, 2013.

[4] V. Dumoulin, E. Perez, N. Schucher, F. Strub, Harm de Vries, A. Courville, and Y. Bengio, "Feature-wise transformations," *Distill*, 2018, https://distill.pub/2018/feature-wise-transformations.

This work is done during Jeju Deep Learning Summer Camp. We also gratefully acknowledge the support from the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and Trompa project.

ICASSP 2019

upf. Music Technology Group

Image Processing Group

EXCELENCIA MARÍA DE MAEZTU

TROMPA