Music Source Separation with Generative Flow

Ge Zhu, Jordan Darefsky, Fei Jiang, Anton Selitskiy, and Zhiyao Duan

Department of Electrical and Computer Engineering, Department of Computer Science, University of Rochester

Abstract

Most existing source separation methods require **full supervision** (i.e., audio mixture and ground-truth sources) for training. In this project, we leverage flow-based generators under **source-only supervision** to learn source priors to separate music mixtures. Experiments show that our proposed method achieves competitive results to one of the full supervision systems, and one variant of our proposed system is capable of separating unseen source tracks.

Music Source Separation

Music source separation aims to isolate a music mixture signal into multiple source signals.

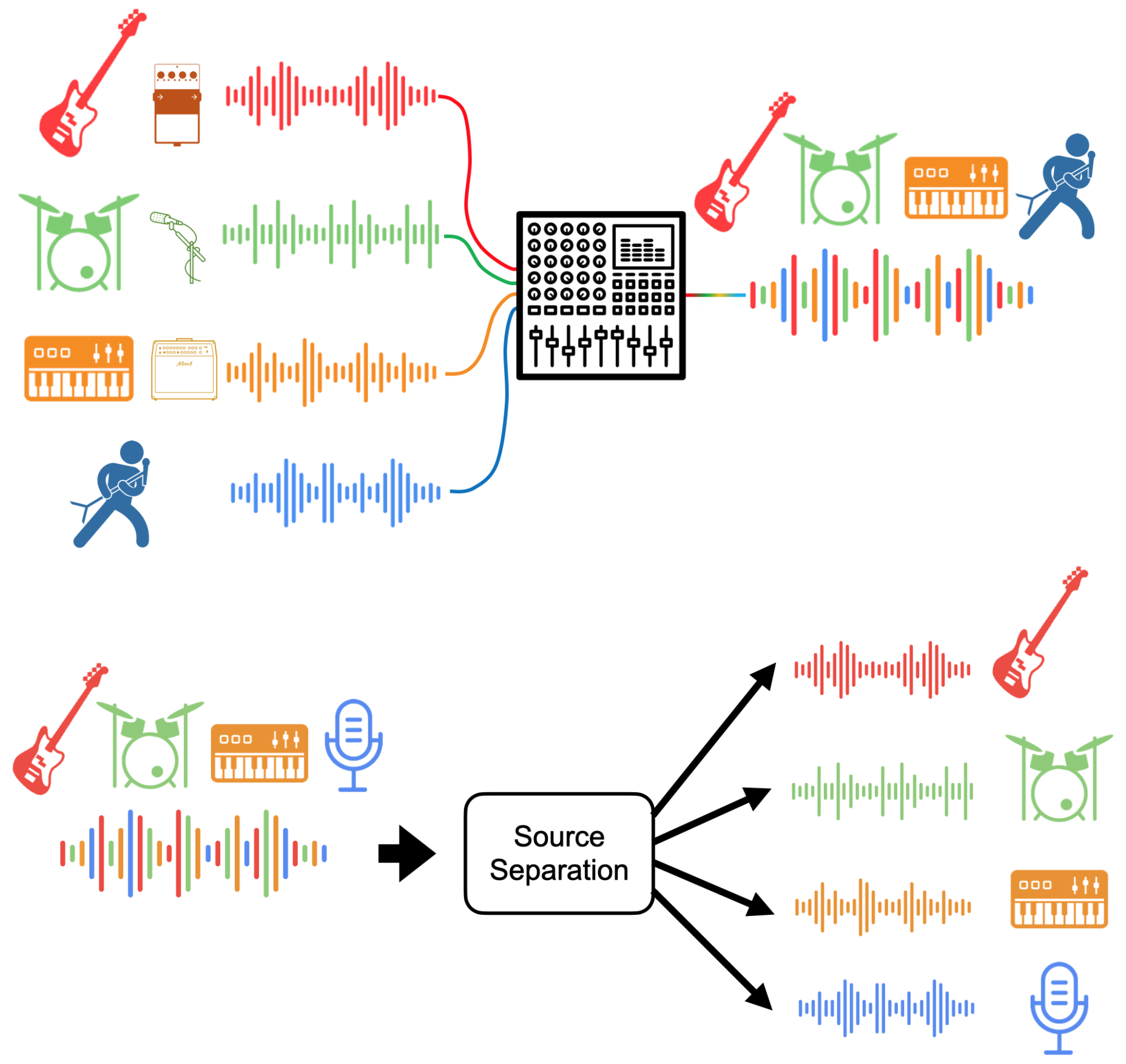


Figure: Music mixing (top) and its inverse process "unmixing" (bottom) or source separation [1].

Source-only supervision

Advantages over full supervision:

- Easier to access isolate and independent instrument tracks;
- Avoid overfitting specific sources combinations;
- More flexible to update new source models without remixing and retraining;
- Analogous to human perception, as we are exposed to either mixtures or isolate sources but not both in parallel.

Background

In generative supervised source separation, a mixture signal $\mathbf{x}$ is composed of several sources $\mathbf{s}_i$ generated by corresponding latent variables $\mathbf{z}_i$:

$$\mathbf{x} = \sum_i^n \mathbf{s}_i, \quad \mathbf{s}_i|\mathbf{z}_i \sim p_G(\mathbf{s}_i|\mathbf{z}_i) \tag{1}$$

where $\mathbf{s}_i$, $i \in \{1, ..., n\}$, indicate different sources.

During training, a group of generative models act as source priors and are trained to approximate source distributions. In traditional probabilistic models, sources are generated through linear or non-linear transformations and the noise distribution $\mathbf{n}$ is assumed to be Poisson distribution, i.e., $p_G(\mathbf{s}_i|\mathbf{z}_i) = \mathcal{PO}(\mathbf{s}_i; f_\theta(\mathbf{z}_i))$. Alternatively, sources can be directly related with latent variables $\mathbf{s}_i = f_i(\mathbf{z}_i)$ using implicit generators.

During inference, sources are extracted by maximizing $p(\mathbf{x}|\sum_i^n \mathbf{s}_i)$ or equivalently minimizing the reconstruction error between the generated and ground-truth mixtures.

Flow-based Generative Source Separation

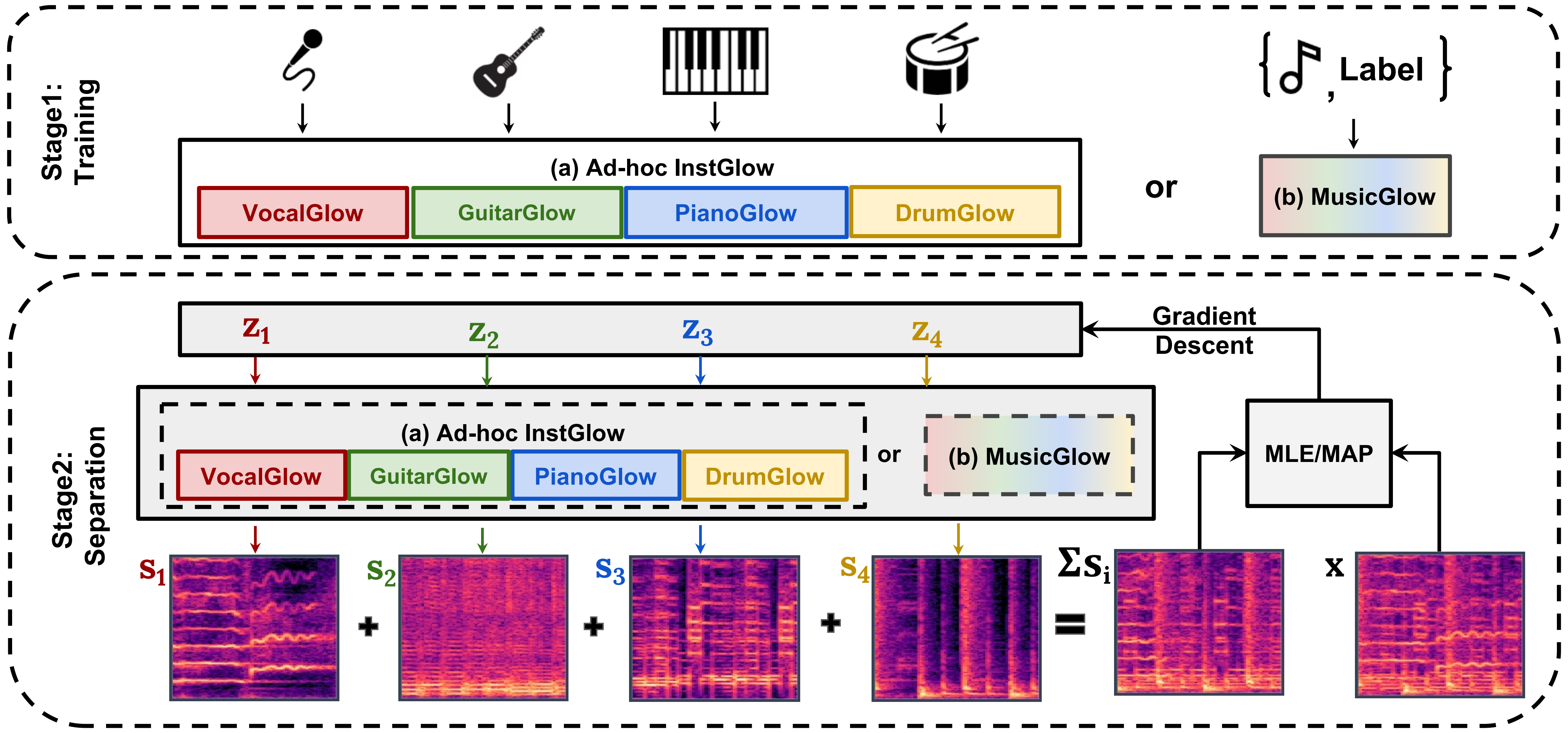
In our method, we use flow-based models as implicit generative priors to model spectrogram. Compared to traditional models, they have the following advantages:

- Invertible, hence it has zero representation error and is capable of recovering any signals;
- Efficient in time and memory for synthesizing signals and inferring latent variables during inference.

During inference, sources $\mathbf{s}_i$ are estimated by searching the optimal latent variables $\mathbf{z}_i$ that maximize the likelihood of the mixture $\mathbf{x}$:

$$\hat{\mathbf{s}}_i = \underset{\mathbf{s}_i=f_i(\mathbf{z}_i)}{\operatorname{argmax}}\ p(\mathbf{x}|\sum_i^n \mathbf{s}_i). \tag{2}$$

It can be seen as the reconstruction error between the mixture $\mathbf{x}$ and the generated mixture $\sum_i^n \mathbf{s}_i$. We apply gradient descent based approach for searching the optimal latent variables. Once we synthesize spectrogram for each component, we use iSTFT to resynthesize audio signal.



Results

We first evaluate full supervision systems and source-only supervision systems on the test partition of MUSDB18 for music source separation. Evaluation metric: global signal-to-distortion ratio (SDR/dB), higher is better.

Method	Vocals	Bass	Drums	Other
instGlow	**3.92**	**2.58**	**3.85**	**2.37**
MusicGlow	2.28	1.27	1.98	1.31
GAN-prior [2]	-0.44	0.48	-0.40	0.32
Demucs (v2) [3]	**7.14**	**5.50**	**6.74**	**4.16**
Wave-U-Net [4]	5.06	2.63	3.74	1.95

We also evaluate on Slakh2100-submix, each track is synthesized from MIDI using professional-grade sample-based virtual instruments. It consists of bass, drums, guitar and piano instrument tracks from its original test set.

Method	Bass	Drums	Guitar	Piano
instGlow	1.54	**6.14**	**1.85**	**0.8**
MusicGlow	**2.57**	5.25	–	–
GAN-prior [2]	0.09	0.85	-0.01	-0.42
Demucs (v2) [3]	**5.48**	**10.21**	–	–
Wave-U-Net [4]	0.01	3.91	–	–

References

[1] E. Manilow, P. Seetharman, and J. Salamon, *Open Source Tools & Data for Music Source Separation*, 2020.

[2] V. Narayanaswamy, J. J. Thiagarajan, R. Anirudh, and A. Spanias, "Unsupervised audio source separation using generative priors," in *Interspeech*, 2020.

[3] A. Défossez, "Hybrid spectrogram and waveform source separation," in *ISMIR*, 2021.

[4] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *ISMIR*, 2018.

Acknowledgements

This work is funded by the New York Center of Excellence in Data Science and gifts from the Voice Biometrics Group.