SVSGAN: SINGING VOICE SEPARATION VIA GENERATIVE ADVERSARIAL NETWORK

Zhe-Cheng Fan, Yen-Lin Lai, Jyh-Shing R. Jang

Department of Computer Science and Information Engineering, National Taiwan University, Taiwan

ABSTRACT

Separating two sources from an audio mixture is an important task with many applications. It is a challenging problem since only one signal channel is available for analysis. In this paper, we propose a novel framework for singing voice separation using the generative adversarial network (GAN) with a time-frequency masking function. The mixture spectra is considered to be a distribution and is mapped to the clean spectra which is also considered a distribution. The approximation of distributions between mixture spectra and clean spectra is performed during the adversarial training process. In contrast with current deep learning approaches for source separation, the parameters of the proposed framework are first initialized in a supervised setting and then optimized by the training procedure of GAN in an unsupervised setting. Experimental results on three datasets (MIR-1K, iKala and DSD100) show that performance can be improved by the proposed framework consisting of conventional networks.

Index Terms— Singing voice separation, music source separation, deep learning, generative adversarial network

1. INTRODUCTION

Monaural source separation is important to various music applications and is sometimes used as a pre-processing step of music signal analysis. For instance, leading instrument detection [1, 2] separates a leading instrument from its accompaniments. Singing pitch estimation [3–5] can be improved by first separating vocals from background music. Cover song identification [6] is also based on leading instrument or vocal pitch features, estimated using a separated singing voice.

Several approaches have been proposed for singing voice separation. Rafii and Pardo proposed the REPET system [7] to separate voice and music by extracting the repeating musical structure. Assumption of low rank and sparsity of music has been used for matrix decompostition [8–11]. The widely used non-negative matrix factorization (NMF) is applied by learning the non-negative reconstruction bases and weights for singing voice separation [12]. Moreover, a complex NMF model [13] has been proposed for jointly estimating the spectrogram and the source phase.

With the development of deep learning, Mass *et al.* [14] used recurrent neural networks (RNN) to create a clean voice.

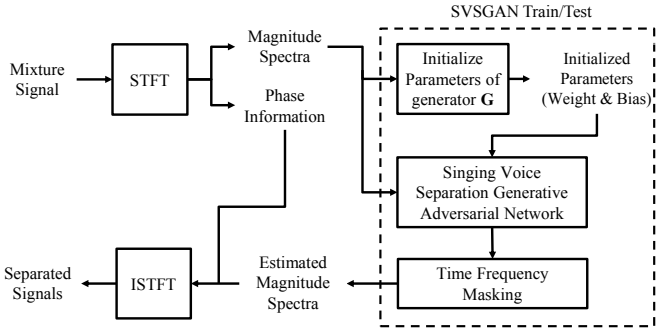

Fig. 1: Block diagram of the proposed framework.

Huang *et al.* [15] then proposed deep RNN with discriminative training to reconstruct vocals from background music. Training multi-context networks [16, 17] with different inputs combined at layer level was proposed to improve audio separation performance. Deep clustering [18] is also used for music separation. Post-processing with a Wiener filter at the output of neural networks and data augmentation [19] have been proposed to separate vocals and instruments. All of these deep learning techniques use multiple non-linear layers to learn the optimal hidden representations from data in a supervised setting.

Generative adversarial networks (GAN) are a new generative model of deep learning [20], which has been successfully used in the field of computer vision to generate realistic images. In the field of source separation, Pascual *et al.* [21] proposed the use of GAN on speech enhancement, which operates in the waveform domain and aims to generate clean vocal waveforms. This paper proposes a novel framework for singing voice separation via GAN (SVSGAN) which operates in the frequency domain and uses a conditional version of GAN. To our knowledge, this is the first proposed framework to use adversarial learning to perform singing voice separation. We regard each spectrum as a sample vector coming from the distribution of spectra. Non-linear mapping of distributions between the mixture spectra and the clean spectra is performed during the adversarial training process. Before adversarial training, the generator parameters are first initialized with joint optimization in a supervised setting and then optimized by the SVSGAN training process in an unsupervised

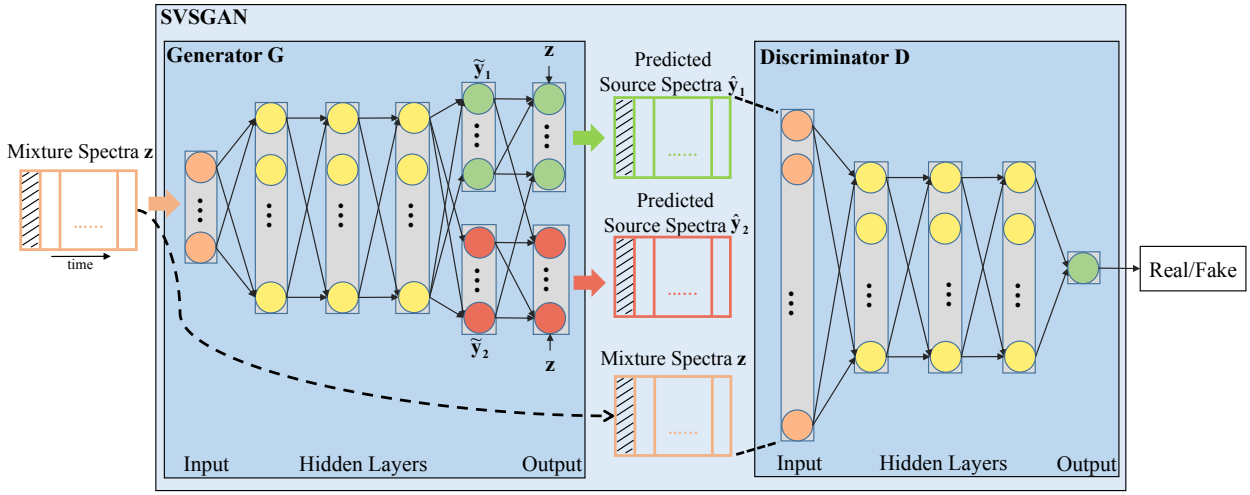

Fig. 2: The proposed SVSGAN framework which consists of two conventional DNNs: generator G and discriminator D. Each spectrum is considered to be a sample vector coming from a distribution of spectra.

setting. Finally the time-frequency masking function consists of the generator outputs. The block diagram of the proposed framework is shown in Fig. 1.

The remainder of this paper is organized as follows: Section 2 gives an overview of GAN. Section 3 presents the details of proposed model including parameter initialization and the adversarial training process. Section 4 presents the experimental settings and results using the MIR-1K, iKala and DSD100 datasets. We conclude the paper in Section 5.

2. GENERATIVE ADVERSARIAL NETWORKS

Ian *et al.* [20] proposed adversarial learning models that learn to map samples z from one distribution to samples x from another distribution. GAN consists of generative model G and discriminative model D, which compete in a two-player min-max game. G aims to imitate the real data distribution while D is a binary classifier which tries to accurately distinguish real data from those generated. Within this min-max game, the generator and the discriminator can be trained jointly by optimizing the following objective function:

$$\min_G \max_D V(G,D) = E_{x\sim P_{data(x)}}[\log D(x)] + E_{z\sim P_{G(z)}}[\log(1 - D(G(z)))], \quad (1)$$

where x is real data sampled from distribution P_{data} and $G(z)$ stands for artificial data sampled from distribution P_G. It is shown in [20] that sufficient training data and epochs allows the distribution P_G to coverage to the distribution P_{data}.

To get more mapping information, we use a conditional extension of GAN (CGAN) [22] which is augmented with some side information. Suppose there is a context vector y as side information, the generator $G(z, y)$ tries to synthesize realistic data under the control of y. Similarly, the CGAN model allows the output of the discriminative model $D(x, y)$ to be controlled by the context vector y. The objective function becomes the following:

$$\min_G \max_D V_{CGAN}(G,D) = E_{x,y\sim P_{data(x,y)}}[\log D(x,y)] + E_{z\sim P_{G(z)},y\sim P_{data}(y)}[\log(1 - D(G(z,y),y))]. \quad (2)$$

In this work, we adjust the input of CGAN, which is discussed in Section 3.

3. PROPOSED WORK

3.1. Model of Singing Voice Separation GAN (SVSGAN)

The SVSGAN architecture consists of two conventional deep neural networks (DNNs): generator G and discriminator D, as shown in Fig. 2. We use magnitude spectra as features and take each spectrum as a sample vector from the spectra distribution. Non-linear mapping is performed between the input mixture spectrum and output clean spectrum, which consists of the vocal part and background music part. Generator G inputs a mixture spectra and generates realistic vocal and background music spectra while discriminator D distinguishes the clean spectra from those generated spectra.

Given that magnitude spectra are transformed from the time domain audio signals using short time Fourier transform (STFT), the output targets $\mathbf{y_1}$ and $\mathbf{y_2}$ of the network are the magnitude spectra of different sources. After training, the network's output predictions, which are also magnitude spectra, are $\tilde{\mathbf{y}}_\mathbf{1}$ and $\tilde{\mathbf{y}}_\mathbf{2}$. The time-frequency masking function, called a soft time-frequency mask, can smooth the source separation results and is used here. The time-frequency mask can be defined as:

$$\mathbf{m}(f) = \frac{|\tilde{\mathbf{y}}_\mathbf{1}(f)|}{|\tilde{\mathbf{y}}_\mathbf{1}(f)| + |\tilde{\mathbf{y}}_\mathbf{2}(f)|}, \quad (3)$$

where $f = 1, 2, ..., F$, stands for different frequencies. After a time-frequency mask is calculated, it is applied to the spectra $\mathbf{z}$ of the mixture signals to estimate the predicted separation spectra $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$, corresponding to source 1 and source 2, defined as:

$$\begin{aligned} \tilde{\mathbf{s}}_1(f) &= \mathbf{m}(f)\mathbf{z}(f), \\ \tilde{\mathbf{s}}_2(f) &= (1 - \mathbf{m}(f))\mathbf{z}(f), \end{aligned} \tag{4}$$

where $f = 1, 2, ..., F$, stands for different frequencies. However, based on [15], the joint optimization is proposed to achieve better results. Similarly, instead of training the network for the time-frequency mask, we train it with the time-frequency masking function. As shown in the left part of Fig.2, the time-frequency masking function is regarded as an additional layer at the network output, defined as:

$$\begin{aligned} \hat{\mathbf{y}}_1 &= \frac{|\tilde{\mathbf{y}}_1|}{|\tilde{\mathbf{y}}_1| + |\tilde{\mathbf{y}}_2|} \otimes \mathbf{z}, \\ \hat{\mathbf{y}}_2 &= \frac{|\tilde{\mathbf{y}}_2|}{|\tilde{\mathbf{y}}_1| + |\tilde{\mathbf{y}}_2|} \otimes \mathbf{z}, \end{aligned} \tag{5}$$

where $\otimes$ stands for element-wise operation. $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ are estimated spectra, which can be transformed into time-domain signals using the inverse short time Fourier transform (ISTFT) with phase information. In this way, the network and time-frequency masking function are jointly optimized. In our proposed framework, the final output separated spectra is based on Eq. 5.

3.2. Training Objective Functions

Before adversarial training, the parameters of the generator G are initialized by performing Eq. 5 in a supervised setting. The training objective J is the mean squared error (MSE) function, which is defined as follows:

$$J = \|\hat{\mathbf{y}}_1 - \mathbf{y}_1\|^2 + \|\hat{\mathbf{y}}_2 - \mathbf{y}_2\|^2. \tag{6}$$

After parameter initialization, the generator G provides basic performance for singing voice separation to serve as the baseline for our experiments.

To fit the input of generator G, the training objective function of SVSGAN by adjusting Eq. 2 is defined as follows:

$$\begin{aligned} &\min_G \max_D V_{SVSGAN}(G, D) = \\ &E_{\mathbf{z},\mathbf{s}_c \sim P_{data(\mathbf{z},\mathbf{s}_c)}}[\log D(\mathbf{s}_c, \mathbf{z})] + \\ &E_{\mathbf{z} \sim P_{G(\mathbf{z})}}[\log(1 - D(G(\mathbf{z}), \mathbf{z})], \end{aligned} \tag{7}$$

where $\mathbf{s}_c$ is the concatenation of $\mathbf{y}_1$ and $\mathbf{y}_2$, and the output of $G(\mathbf{z})$ is the predicted spectra consisting of the concatenation of $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$, which is generated from input spectra $\mathbf{z}$. The output of discriminator D is controlled by the augmented input spectra $\mathbf{z}$. By this step, the SVSGAN not only approximates the distribution between input spectra and output spectra but also learns the general structure of the spectra.

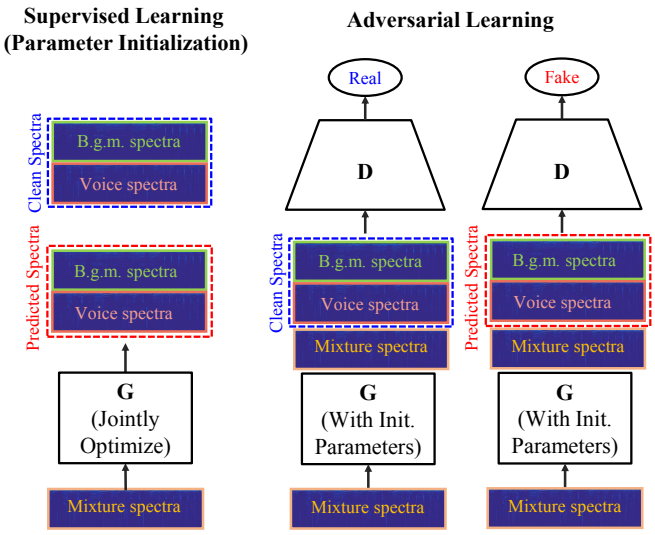

Fig. 3: SVSGAN training process, where "B.g.m." stands for background music. The parameters of generator G are first initialized. Discriminator D returns "fake" when the input contains predicted spectra and returns "real" when input contains clean spectra.

In addition, we use log D trick [20] as the objective function for generator G.

Note that better separation results may be obtained using complicated training objective functions and more powerful neural networks, such as RNN or CNN. However, we use a basic neural network architecture and the MSE as the training objective to investigate the degree of performance improvement provided by GAN.

4. EXPERIMENTS

4.1. Dataset & Settings

The proposed framework is evaluated using the MIR-1K dataset [23], iKala dataset [11] and Demixing Secret Database (DSD100) [24]. The MIR-1K dataset consists of 1,000 song clips lasting 4 to 13 seconds with a sample rate of 16,000 Hz. These clips are recorded from 110 Chinese popular karaoke songs performed by both male and female amateurs. The iKala dataset consists of 352 30-second song clips with a sample rate of 44,100 Hz. These clips are recorded from Chinese popular songs performed by professional singers. Only 252 song clips are released as a public subset for evaluation. Each song clip in these two datasets is a stereo recording, with one channel for the singing voice and the other for background music. Manual annotations of the pitch contours are provided. In experimental settings, we randomly select one-fourth of the song clips for training data and the remaining song clips are used for testing.

The DSD100 dataset is taken from a subtask called MUS from the Signal Separation Evaluation Campaign (SiSEC). It

MIR-1K Dataset			
Model	SDR	SAR	SIR
DNN (baseline)	6.57	10.14	9.84
SVSGAN (V+B)	6.69	10.32	9.86
SVSGAN (V+M)	6.73	10.28	9.96
SVSGAN (V+B+M)	6.78	10.29	10.07
IBM (upper bound)	13.92	14.80	21.96
iKala Dataset			
Model	SDR	SAR	SIR
DNN (baseline)	9.74	11.72	14.99
SVSGAN (V+B)	10.15	12.48	14.72
SVSGAN (V+M)	10.22	12.78	14.41
SVSGAN (V+B+M)	10.32	12.87	14.54
IBM (upper bound)	12.30	14.10	23.70

Table 1: Vocal results (in dB) of conventional DNN and SVSGANs on the MIR-1K and iKala datasets. "IBM" represents ideal binary mask. Some examples of singing voice separation are provided at `http://mirlab.org/demo/svsgan`.

consists of Dev and Test parts each with 50 songs with a sample rate of 44,100 Hz. Each song provides four sources: bass, drums, other and vocals and the mixture is semi-professionally engineered. The average duration of these songs is 4 minutes and 10 seconds and the dataset includes a wide variety of music genres.

To reduce computational cost, all song clips from the iKala and DSD100 datasets are downsampled to 22,050 Hz. We used STFT to yield magnitude spectra with a 1024-point window size and a 256-point hop size. Performance is measured in terms of source to distortion ratio (SDR), source to interference ratio (SIR), and source to artifact ratio (SAR), calculated by the Blind Source Separation (BSS) Eval toolbox v3.0 [25]. For the iKala and MIR-1K datasets, overall performance is reported on weighted means of the SDR, SAR and SIR. For the DSD100 dataset, overall performance is reported on median values of SDR based Test part.

4.2. Experimental Results

To compare the performance between the conventional DNN and SVSGANs, we construct a conventional DNN, which consists of 3 hidden layers, each with 1024 neurons, denoted as DNN (baseline). The architecture of generator G in the SVSGANs is identical to the baseline and is combined with discriminator D consisting of 3 hidden layers, each with 512 neurons. The difference between the SVSGANs is the input spectra of discriminator D, as shown in Table 1, where "V" stands for the vocal spectra, "B" is the background music spectra, and "M" is the mixture spectra. Comparing DNN (baseline) to SVSGANs, the results on iKala and MIR-1K datasets show that SVSGANs enhance performance in terms of SDR and SAR. Comparing different SVSGAN architectures, SVSGAN (V+B) represents the results of the original

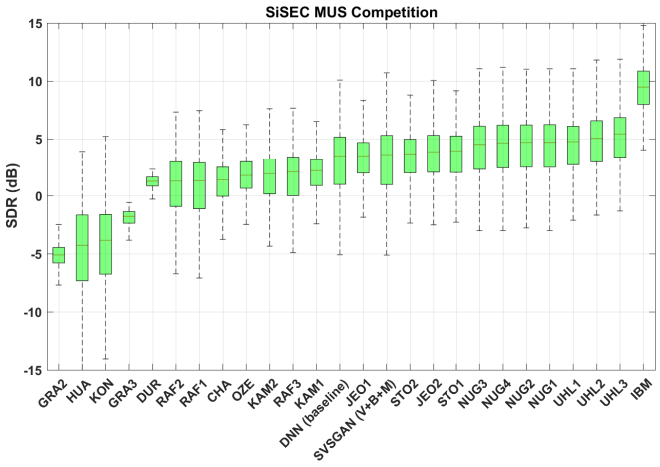

Fig. 4: Vocal results on the Test part of the DSD100 dataset, sorted by median values of each submission.

GAN architecture while SVSGAN (V+M) and SVSGAN (V+B+M) represent the results of the conditional GAN. SVSGAN (V+M) is found to provide better results, indicating that when the input of discriminator D contains the mixture spectra, SVSGAN (V+M) not only learns the mapping from the distribution of mixture spectra to the distribution of clean spectra but also learns a general structure from the mixture spectra at the same time. Comparing SVSGAN (V+M) to SVSGAN (V+B+M), which has more inputs for discriminator, suggests that increasing the number of inputs to discriminator D improves performance.

Fig. 4 compares the Test part of the DSD100 dataset. The DNN (baseline) and SVSGAN (V+B+M) are the same as those evaluated on the iKala and MIR-1K datasets. Since we only trained the model with the Dev part of dataset without additional augmented datasets, such as MedleyDB [26], SVSGAN (V+B+M) does not outperform all other submissions. However, the result still shows that singing voice separation can be improved by adversarial learning on this dataset featuring a wide variety of music genres.

5. CONCLUSIONS & FUTURE WORK

This paper proposes a singing voice separation model with time-frequency masking function for monaural recordings using a generative adversarial framework. The framework consists of two conventional neural networks with conditional GAN, and is shown to potentially enhance source separation performance. Possible future work involves three directions. First, we will incorporate additional augmented data in our adversarial training process to achieve better performance. Next, we will seek to improve generator G and discriminator D using more powerful neural networks, such as CNN, RNN and other complicated architectures. Finally, we will explore the use of Wasserstein GAN [27] to achieve better performance. Future work will also include further comparisons between SVSGANs and other competitive approaches.

6. REFERENCES

[1] J. L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1180–1191, Oct 2011.

[2] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2015, pp. 2135–2139.

[3] C. L. Hsu, D. Wang, J.-S. R. Jang, and K. Hu, "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," *IEEE/ACM Trans. Audio, Speech, Language Processing*, vol. 20, no. 5, pp. 1482–1491, July 2012.

[4] Z.-C. Fan, J.-S. R. Jang, and C.-L. Lu, "Singing voice separation and pitch extraction from monaural polyphonic audio music via DNN and adaptive pitch tracking," in *Proc. IEEE Int. Conf. Multimedia Big Data. (BigMM)*, 2016, pp. 178–185.

[5] Y. Ikemiya, K. Itoyama, and K. Yoshii, "Singing voice separation and vocal f0 estimation based on mutual combination of robust principal component analysis and subharmonic summation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2084–2095, Nov 2016.

[6] J. Serrà, E. Gómez, and P. Herrera, *Audio Cover Song Identification and Similarity: Background, Approaches, Evaluation, and Beyond*, pp. 307–332, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[7] Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 1, pp. 73–84, Jan 2013.

[8] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2012, pp. 57–60.

[9] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Real-time online singing voice separation from monaural recordings using robust low-rank modeling," in *Proc. Int. Soc. Music Info. Retrieval Conf. (ISMIR)*, 2012, pp. 67–72.

[10] Y.-H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *Proc. Int. Soc. Music Info. Retrieval Conf. (ISMIR)*, 2013, pp. 427–432.

[11] T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and R. Jang, "Vocal activity informed singing voice separation with the iKala dataset," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2015, pp. 718–722.

[12] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," *IEEE/ACM Trans. Audio, Speech, Language Processing*, vol. 21, no. 10, pp. 2096–2107, Oct 2013.

[13] P. Magron, R. Badeau, and B. David, "Complex NMF under phase constraints based on signal modeling: Application to audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2016, pp. 46–50.

[14] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. Interspeech*, 2012, pp. 22–25.

[15] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proc. Int. Soc. Music Info. Retrieval Conf. (ISMIR)*, 2014, pp. 477–482.

[16] X.-L. Zhang and D. Wang, "Multi-resolution stacking for speech separation based on boosted dnn," in *Proc. Interspeech*, 2015, pp. 1745–1749.

[17] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 967–977, Mar 2016.

[18] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2017, pp. 61–65.

[19] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Imporving music source separation based on deep neural networs through data augmentation and network blenging," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2017, pp. 261–265.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[21] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017.

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *arXiv:1611.07004*, 2016.

[23] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 310–319, Feb 2010.

[24] "SiSEC MUS Homepage," 2016, [Online] https://sisec.inria.fr/sisec-2016/2016-professionally-produced-music-recordings/.

[25] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[26] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive mir research," in *Proc. Int. Soc. Music Info. Retrieval Conf. (ISMIR)*, 2014, pp. 155–160.

[27] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," in *arXiv:1701.07875*, 2017.