Hierarchic Temporal Convolutional Network With Cross-Domain Encoder for Music Source Separation

Article *in* Signal Processing Letters, IEEE · January 2022

DOI: 10.1109/LSP.2022.3187316

5 authors, including:

Ying Hu
Xinjiang University
22 PUBLICATIONS **76** CITATIONS
SEE PROFILE

Yadong Chen
Xinjiang University
2 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Liang He
Tsinghua University
92 PUBLICATIONS **676** CITATIONS
SEE PROFILE

Hao Huang
Xinjiang University
53 PUBLICATIONS **261** CITATIONS
SEE PROFILE

Hierarchic Temporal Convolutional Network With Cross-Domain Encoder for Music Source Separation

Ying Hu, Yadong Chen, Wenzhong Yang, Liang He, and Hao Huang

***Abstract*—Recently, the time-domain-based methods (i.e., the method of modeling the raw waveform directly) for audio source separation have shown tremendous potential. In this paper, we propose a model which combines the complexed spectrogram domain feature and time-domain feature by a cross-domain encoder (CDE) and adopts the hierarchic temporal convolutional network (HTCN) for multiple music sources separation. The CDE is designed to enable the network to code the interactive information of the time-domain and complexed spectrogram domain features. HTCN enables it to learn the long-time series dependence effectively. We also designed a feature calibration unit (FCU) to be applied in the HTCN and adopted the multi-stage training strategy during the training stage. The ablation study demonstrates the effectiveness of each designed component in the model. We conducted the experiments on the MUSDB18 dataset. The experimental results indicate that our proposed CDE-HTCN model outperforms the top-of-the-line methods and, compared with the state-of-the-art method, DEMUCS, achieves the improvement of the average SDR score of 0.61 dB. Significantly, the improvement of the SDR score for the $bass$ source has a sizable margin of 0.91 dB.**

***Index Terms*—Music source separation, time-domain, cross-domain feature, temporal convolutional network.**

I. Introduction

MUSIC source separation is one of the fundamental research areas in music signal processing. It has many use cases, such as generating the accompaniment from pop songs for automatic karaoke, separating the specific sources as a preprocessing tool [1] for other tasks such as music transcription [2] and DJ-related applications [3].

In recent years, neural network-based methods have obtained remarkable results for music source separation [4], [6], [7], [10], [12]. These methods, often trained in a supervised setting, can be divided into two categories: the methods based on the complexed spectrogram domain [7], [8], [12] and on the time domain [6], [10], [15]. The former utilizes a time-frequency representation produced by a short-time Fourier transform (STFT). The latter directly models the raw audio waveform and has achieved competitive performance. Most methods based on the time domain transform the input signal into a learnable latent space and generate the masks of target sources, then further reconstruct the corresponding waveform by exploiting the mixture signal. TasNet [18], a notable speaker separation approach based on the time domain, generates the masks exploiting a learnable front-end obtained from an LSTM and yields comparable performance. TasNet consists of three parts: encoder, separator, and decoder. The encoder can be considered a learnable STFT operation, and the decoder a learnable inverse STFT operation [17]. Then, Luo and Mesgarani [16] proposed Conv-TasNet, the LSTM in TasNet replaced with a superposition of dilated CNNs, which further improved the separation performance. Recently, David *et al.* designed a hierarchical meta-learning-inspired model [6], Meta-TasNet, which successfully modified the Conv-TasNet architecture for music source separation. Alexandre *et al.* designed an encoder-decoder model with a U-Net structure and bidirectional LSTM [10], DEMUCS, which is more expressive than the Conv-TasNet. Typically, music relies heavily on its repetitions to build a logical structure. Although the time domain waveform can partly describe the various realizations of a sound, many repetitive features can be readily available from the spectrogram [21]. In addition, models such as Conv-TasNet are typically built with a pure CNN-based encoder/decoder architecture, and it might fail to perform well when only a small-scale dataset is made available for training its component models, probably leading to incorrect estimation of the distribution for audio features [23]. Building on these observations, we propose a superior framework with a **c**ross-**d**omain **e**ncoder and **h**ierarchic **t**emporal **c**onvolutional **n**etwork (CDE-HTCN) to improve the performance of multiple music sources separation. CDE-HTCN comprises an encoder-separator-decoder architecture with a similar fashion as Conv-TasNet. The main contributions of this letter are as follows:

i) We propose a cross-domain encoder structure to leverage the features of two domains derived from a 1-D convolution layer and an STFT operation. We also explore a novel feature fusion scheme.

ii) We propose a more powerful separator named hierarchic temporal convolutional network (HTCN) and design a feature calibration unit (FCU) to be equipped in the HTCN.

Manuscript received 14 April 2022; revised 13 June 2022; accepted 17 June 2022. Date of publication 30 June 2022; date of current version 13 July 2022. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U1903213 and in part by the Tianshan Innovation Team Plan Project of Xinjiang under Grant 202101642. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Maximo Cobos. *(Yadong Chen and Ying Hu contributed equally to this work.) (Corresponding author: Ying Hu.)*

Ying Hu, Yadong Chen, Wenzhong Yang, and Hao Huang are with the College of information science and engineering, Xinjiang University, Urumqi 830000, China (e-mail: huying@xju.edu.cn; cyd@stu.xju.edu.cn; ywz_xy@163.com; huanghao@xju.edu.cn).

Liang He is with the Department of electrical engineering, Tsinghua University, Beijing 100084, China, and also with the College of information science and engineering, Xinjiang University, Urumqi 830000, China (e-mail: heliang@mail.tsinghua.edu.cn).

Digital Object Identifier 10.1109/LSP.2022.3187316

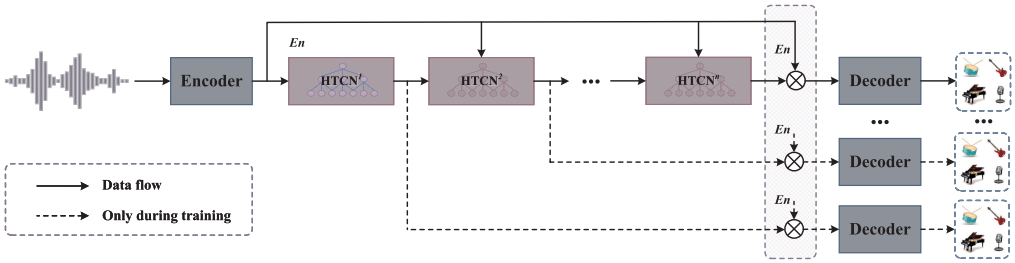


Fig. 1. Diagram of the CDE-HTCN architecture. The dash lines represent the procedures that exist only during training. En represents the output after encoding and $\otimes$ the element-wise multiplication.

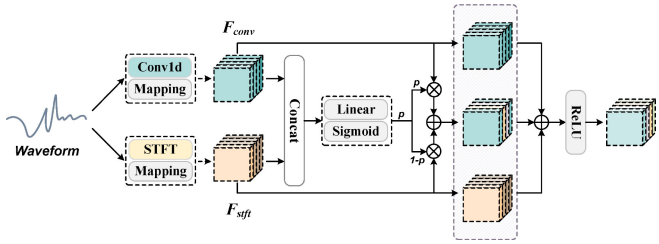


Fig. 2. Diagram of the cross-domain encoder. $\oplus$ represents the element-wise addition and $\otimes$ the element-wise multiplication. p is a weighted factor.

The rest of this letter is organized as follows. Our proposed method will be introduced in Section II and the experimental results presented in Section III, then a conclusion drawn in Section IV.

II. Algorithm Description

The illustration of our proposed CDE-HTCN is shown in Fig. 1. The network consists of an encoder for feature generation, serial HTCNs for music-related feature extraction, and multiple decoders in multiple stages, each of which reconstructs the waveforms of multiple target sources. Similar to [19] and [20], the CDE-HTCN model adopts a multi-stage training strategy, which necessitates outputting the waveforms of each target source at the end of each decoder.

A. Cross-Domain Encoder (CDE)

The time-domain features obtained by a 1-D CNN and complexed spectragram by the STFT operation are processed in parallel by the cross-domain encoder. To effectively select the features from across domains, we develop a fusion scheme that combines each other's features to preserve key cues and discard inessential information. This is conceptually similar to the recent attempts in the computer vision community [25] and automatic speech recognition field [23].

In detail, the waveform of music source *c* denotes $x_c \in \mathbb{R}^{2\times T}$ and that of stereo mixture $x \in \mathbb{R}^{2\times T}$, where *2* is the number of channels (for stereo) and *T* that of samples. As shown in Fig. 2, the time domain feature F_{conv} and complexed spectrogram domain feature F_{stft} are derived in a parallel manner by a 1-D CNN and STFT operation, respectively. $F_{stft} \in \mathbb{R}^{2F\times L}$ is formed by the concatenation of the real and imaginary parts of complexed spectragram along with the frequency dimension, where L denotes the number of time frames and F that of frequency bins. F_{conv} and F_{stft} are further transformed to keep with the same dimension of feature representations by a linear layer, respectively. Subsequently, after a concatenation operation, a linear layer and *sigmoid* activation operation denoted as σ are exploited to generate a gated parameter p for F_{conv} and thus *1-p* for F_{stft}. The whole procedure can be formulated as:

$$p = \sigma(w(concat[F_{conv}, F_{stft}])) \tag{1}$$

$$En = \Theta((p \odot F_{conv} + (1-p) \odot F_{stft}) + F_{conv} + F_{stft}) \tag{2}$$

where w denotes the linear operation and $\odot$ the element-wise multiplication, Θ the rectified linear unit (ReLU) activation operation. The output of cross-domain encoder denotes as $En \in \mathbb{R}^{N\times L}$, where *N* is the number of channels and *L* that of time frames.

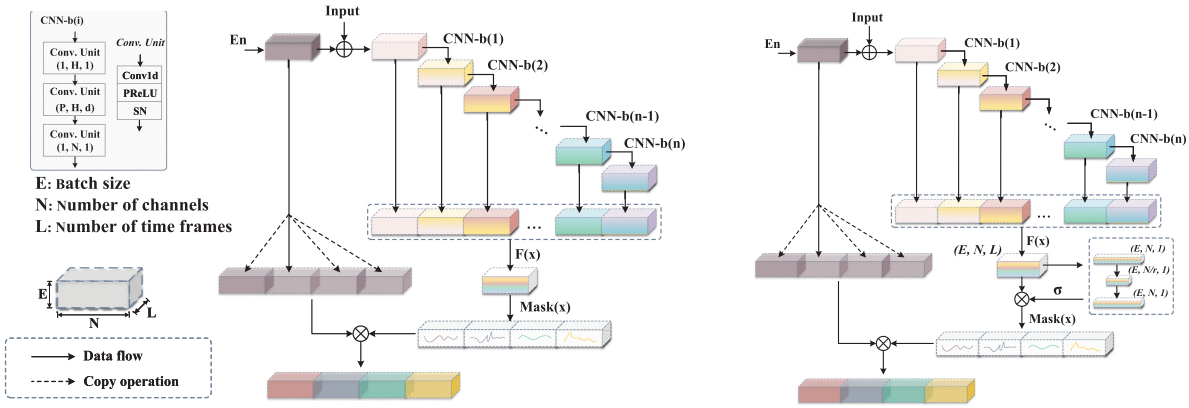
B. Hierarchic Temporal Convolutional Network (HTCN)

Based on the temporal convolutional network (TCN) in Conv-TasNet [16], we propose an HTCN consisting of cascade CNN blocks and can be used in multiple stages. As shown in Fig. 3(a), the dilation factors d of cascade CNN blocks are exponentially increasing, which ensures the temporal context window is sufficiently larger to take advantage of the long-range dependencies of the music signal. Each CNN layer is followed with a PReLU activation and switchable normalization (SN) operations [27], which is the weighted sum of instance normalization (IN), batch normalization (BN), and layer normalization (LN). The weights are learned by training the network that has been proved to be more robust for tasks.

The correlation between the time sequences is crucial for the audio separation task. We designed the HTCN with a one-shot aggregation mechanism [28] to aggregate the time sequence features effectively obtained by the exponential increasing dilation rate. As shown in Fig. 3, a linear layer $F(\cdot)$ integrates the concatenated features x linearly and compress the number of channels. The output of $F(x)$ is further used for mask estimation, which consists of a 1-D CNN and nonlinear activation layers, and outputs C estimated masks matrices $\hat{m}_i \in \mathbb{R}^{N\times L}, i = 1, \ldots, C$. The whole HTCN can be formulated as:

$$\hat{m}_i = Mask(F(concat[f_1, f_2, \ldots, f_n])) \tag{3}$$

where f_n denotes the output of the CNN-b(n), $F(\cdot)$ the linear operation and $Mask(\cdot)$ the masking operation.



(a) HTCN (b) HTCN with Feature Calibration Unit

Fig. 3. The HTCN estimates the C masks of music sources here $C = 4$. d denotes a dilation factor of CNN in the CNN-b(n) block. (a): Diagram of the HTCN. (b): Diagram of the HTCN with the feature calibration unit. $F(\cdot)$ denotes the linear operation and $Mask(\cdot)$ the masking operation. The CNN block and a nonlinear activation function estimate C masks.

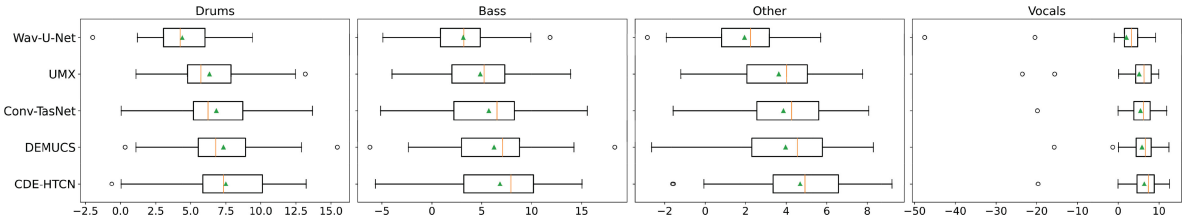


Fig. 4. Boxplots of SDR for different models on all target sources. Note that distributional outliers, i.e., data lying outside the 99.3 % distribution, are represented by small "o" signs in the plots for readability.

Feature Calibration Unit: The hierarchical structure using a one-shot connection mode can optimize gradient propagation and aggregate richer shallow features. However, it is not sufficient to integrate the features from all the previous layers only through a simple linear operation, which may cause the semantic information of deep layers to be blurred and the subsequent mask estimation to be inaccurate. To improve the structure of the network and boost the separation performance, we design a channel attentive feature descriptor, called feature calibration unit (FCU), as shown in Fig. 3(b).

Similar to the Squeeze-Excitation network [29], the FCU compresses the spatial dependency by adaptive average pooling to learn a channel-specific descriptor and then rescales the input feature map for highlighting useful channels through two linear layers followed by a hard-swish function. The recalibrated features F_{fcu} can be described as:

$$F_{fcu} = f' \odot hswish(w_2(\delta(w_1(f_{Aap}(f'))))) \tag{4}$$

where f_{Aap} denotes the operation of adaptive average pooling, $w_1 \in \mathbb{R}^{\frac{N}{r} \times N}$ and $w_2 \in \mathbb{R}^{N \times \frac{N}{r}}$ the weights parameters of two fully connected layers, respectively, i.e. a dimensionality-reduction layer with parameters w_1 and reduction ratio r (this parameter is set to 4 in this letter), and then a dimensionality increasing layer with parameters w_2. δ denotes $ReLU6$ and $hswish$ the hard-swish activation functions.

C. Decoder

The decoder reconstructs the waveforms of separated target music sources from high-dimensional feature space by a 1-D deconvolutional layer. Although the encoder inputs are cross-domain features, experiments in [21] show that compared with reconstructing from the time domain, reconstructing waveforms from the cross domains only makes a minor difference. Therefore, for the decoder, we adopt the decoding structure in Conv-TasNet: a high-dimensional feature is transformed by a 1-D deconvolutional layer, whose kernel has the same size as that of the convolutional layer in the encoder. Finally, the decoder outputs the C separated music source waveforms.

D. Multi-Stage Training Strategy

During the training process, a multi-scale loss is used to calculate the weighted summation of the loss of each stage [19], [20]. It requires reconstructing the waveforms of C estimated source at each stage. In the testing process, only the last stage outputs the C waveforms of the estimated source. Let x denotes the waveform of mixture song, and y_i that of the i-th target source. The estimated mask of the i-th source in the j-th stage of separation network is $\hat{m}_{i,j}$. The corresponding loss function of the j-th stage for the i-th source is defined as:

$$\mathcal{L}_{i,j} = \|y_i - x \odot \hat{m}_{i,j}\|_{1,1} \tag{5}$$

TABLE I
THE ABLATION STUDY ON MUSDB18 DATASET

Models	Bass	Drums	Other	Vocals	AVG.
Conv-TasNet	6.53	6.23	4.26	6.21	5.81
w/ MS (**i**)	6.88	6.60	4.31	6.54	6.08
w/ HTCN (**ii**)	7.08	6.70	4.54	7.01	6.33
Baseline (**iii**)	7.44	7.07	4.73	7.14	6.60
+ CDE (**iv**)	7.74	7.20	**5.00**	7.35	6.82
+ FCU (**v**)	7.77	**7.35**	4.99	7.18	6.82
+ CDE + FCU (**vi**)	**7.92**	7.33	4.92	**7.37**	**6.89**

The significance of bold entities indicate best values.

Deep supervision is adopted that each stage is assigned to a different weight for the loss. The loss function is defined as:

$$\mathcal{L} = \sum_{j}^{S} \left(w_j \cdot \sum_{i}^{C} \mathcal{L}_{i,j} \right) \tag{6}$$

where S is the number of stages and w_j the weight of j-th stage. Empirically, we set w_j to 1.

III. EXPERIMENTS

A. Experimental Setup

We use the MUSDB18 dataset [30] which consists of 86 trains, 14 validation, and 50 test tracks with full supervision in stereo and sampled at 44.1 kHz. Each song has the exact waveforms of each of the sources, such as the $bass$, $drums$, $other$, and $vocals$ parts. We trained the models for 300 epochs on 4-second long segments. We used the Adam [31] optimizer with a learning rate of 3e-4. The filter length in the encoder and decoder is set to 2 ms and batch size 4. The detailed parameter settings and codes implementation are available online.[1]

B. Experimental Results

We conducted an ablation study to verify the effectiveness of each component in the proposed network. Several variants are compared in Table I: (i) The multi-stage training strategy (MS) is applied to the vanilla Conv-TasNet; (ii) The TCN in the vanilla Conv-TasNet is replaced with HTCN; (iii) The model adopting both the MS and HTCN based on the Conv-TasNet is considered as the baseline model; (iv) The cross-domain encoder (CDE) is applied to the baseline; (v) The feature calibration unit (FCU) is added in HTCN of baseline; (vi) Both CDE and FCU are applied to the baseline.

We observe that each modification over the vanilla Conv-TasNet improves the performance. Compared with the Conv-TasNet, the model replacing the TCN with the HTCN results in the most significant gains (0.52 dB increase of the average SDR on four sources), and the baseline model applying the multi-stage training strategy and HTCN improves the average SDR by 0.79 dB. Furthermore, based on the baseline model, our designed CDE and FCU improve the performance with the average SDR gains of 0.22 dB, respectively. The results show that the cross-domain features obtained by combining spectrum information at the encoder can effectively improve the upper limit of separation performance. Moreover, the results verify that the feature recalibration idea discussed in Section II-B positively impacts separation performance.

TABLE II
COMPARISON OF THE MEDIAN SDR (DB) WITH THE STATE-OF-THE-ART MODELS ON MUSDB18

Models	Param.	Bass	Drums	Other	Vocals	AVG.
IRM oracle	N/A	*7.12*	*8.45*	*7.85*	*9.43*	*8.21*
Wave-U-Net* [15]	10.2 M	3.21	4.22	2.25	3.25	3.23
UMX [8]	8.9 M	5.23	5.73	4.02	6.32	5.33
Meta-TasNet* [6]	45.5 M	5.58	5.91	4.19	6.40	5.52
MMDenseLSTM [11]	4.9 M	5.16	6.41	4.15	6.60	5.58
Sams-Net [5]	3.7 M	5.25	6.63	4.09	6.61	5.65
X-UMX [26]	9.5 M	5.43	6.47	4.64	6.61	5.79
Conv-TasNet* [16]	8.9 M	6.53	6.23	4.26	6.21	5.81
LaSAFT [7]	N/A	5.63	5.68	4.87	7.33	5.88
Spleeter [9]	39.3 M	5.51	6.71	4.02	6.86	5.91
D3Net [12]	7.9 M	5.25	7.01	4.53	7.24	6.01
DEMUCS* [10]	648 M	7.01	6.86	4.42	6.84	6.28
CDE-HTCN*	12.6 M	**7.92**	**7.33**	**4.92**	**7.37**	**6.89**

*indicates the method modeling directly in the time domain
The significance of bold entities indicate best values.

We also compare the proposed network (**CDE-HTCN**) with the state-of-the-art methods. As shown in Table II, the median SDR scores here are either taken from the published papers or SiSEC18 evaluation [32]. It is worth mentioning that no extra data is used in our training procedure. In order to make a fair comparison, we only compare the methods without data augmentation. We also provide the metrics of the Ideal Ratio Mask oracle (IRM), which computes the best possible mask using the ground truth spectrogram. As can be seen from the comparison, our proposed CDE-HTCN outperforms the state-of-the-art method [10] 0.61 dB average SDR score in the separation performance of four music sources and has a sizable margin in $bass$ (0.91 dB higher than DEMUCS [10]).

We also provide a statistical description of the separation performance shown as boxplots for different models on all target sources. As seen in Fig. 4, the separation results of our model show a slightly higher degree of dispersion (relative to other models). In the case of outliers, the separation results of our model are distributed in a high-value range, which shows that the model has good robustness. In addition, we provide additional comparative experiments[2] and some separation demos.[2]

IV. CONCLUSION

In this paper, we propose CDE-HTCN for music source separation. The results show the effectiveness of the multi-stage training strategy and designed HTCN. It verifies that the CDE can improve the separation performance by focusing on the cross-domain feature interactions. Our designed feature calibration unit applied in the HTCN can effectively recalibrate the features before masking estimation. The experimental results show that our proposed CDE-HTCN outperforms the state-of-the-art methods.

[1][Online]. Available: https://github.com/YadongChen-1016/Music-Source-Separation-master

[2][Online]. Available: http://yadongchen-1016.github.io

REFERENCES

[1] Z. Zafar, A. Liutkus, F. R. Stoter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 8, pp. 1307–1335, Aug. 2018.

[2] W. Zhang, Y. Zhang, Y. She, and J. Shao, "Stereo feature enhancement and temporal information extraction network for automatic music transcription," *IEEE Signal Process. Lett.*, vol. 28, pp. 1500–1504, 2021.

[3] L. Veire Vande and T. De Bie, "From raw audio to a seamless mix: Creating an automated DJ system for drum and bass," *EURASIP J. Audio, Speech, Music Process.*, vol. 13, pp. 1–21, 2018.

[4] O. Slizovskaia, G. Haro, and E. Gómez, "Conditioned source separation for musical instrument performances," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2083–2095, 2021.

[5] T. Li, J. Chen, H. Hou, and M. Li, "Sams-Net: A sliced attention-based neural network for music source separation," in *Proc. 12th Int. Symp. Chin. Spoken Lang. Process.*, 2021, pp. 1–5.

[6] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning extractors for music source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 816–820.

[7] W. Choi, M. Kim, J. Chung, and S. Jung, "LaSAFT: Latent source attentive frequency transformation for conditioned source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 171–175.

[8] F. R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix-A reference implementation for music source separation," *J. Open Source Softw.*, vol. 4 no. 41, 2019, Art. no. 1667.

[9] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and efficient music source separation tool with pre-trained models," *J. Open Source Softw*., vol. 5 no. 50, 2020, Art. no. 2154.

[10] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," 2021, *arXiv:1911.13254*.

[11] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement*, 2018, pp. 106–110.

[12] N. Takahashi and Y. Mitsufuji, "D3Net: Densely connected multidilated DenseNet for music source separation," 2020, *arXiv:2010.01733*.

[13] S. Park, T. Kim, K. Lee, and N. Kwak, "Music source separation using stacked hourglass networks," in *Proc. 19th Int. Soc. Music Inf. Retrieval*., 2018, pp. 289–296.

[14] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band DenseNets for audio source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 21–25.

[15] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th Int. Soc. Music Inf. Retrieval*, 2018, pp. 334–340.

[16] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[17] C. Tang *et al.*, "Joint time-frequency and time domain learning for speech enhancement," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 3816–3822.

[18] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*., 2018, pp. 696–700.

[19] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. Int. Conf. Mach. Learn.*, 2020, vol. 119, pp. 7164–7175.

[20] K. Tan, B. Xu, A. Kumar, E. Nachmani, and Y. Adi, "SAGRNN: Self-attentive gated RNN for binaural speaker separation with interaural cue preservation," *IEEE Signal Process. Lett.*, vol. 28, pp. 26–30, 2021.

[21] G. P. Yang, C. I. Tuan, H. Y. Lee, and L. S. Lee, "Improved speech separation with time-and-frequency cross-domain joint embedding and clustering," in *Proc. Interspeech*, 2019, pp. 1363–1367.

[22] T. Lan, Y. Qian, Y. Lyu, R. Mokhosi, W. Tai, and Q. Liu, "Improved speech separation with time-and-frequency cross-domain feature selection," in *Proc. Interspeech*, 2021, pp. 3525–3529.

[23] F.-A. Chao, J.-W. Hung, and B. Chen, "Cross-domain single-channel speech enhancement model with BI-Projection fusion module for noise-robust ASR," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[24] Y. Koyama, T. Vuong, S. Uhlich, and B. Raj, "Exploring the best loss function for DNN-based low-latency speech enhancement with temporal convolutional networks," 2020, *arXiv:2005.11611*.

[25] F. E. Wang, Y. H. Yeh, M. Sun, W. C. Chiu, and Y. H. Tsai, "BiFuse: Monocular 360 depth estimation via bi-projection fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 462–471.

[26] R. Sawata, S. Uhlich, S. Takahashi, and Y. Mitsufuji, "All for one and one for all: Improving music separation by bridging networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*., 2021, pp. 51–55.

[27] P. Luo, R. Zhang, J. Ren, Z. Peng, and J. Li, "Switchable normalization for learning-to-normalize deep representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 712–728, Feb. 2021, doi: 10.1109/TPAMI.2019.2932062.

[28] Y. Lee, J. W. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and GPU-computation efficient backbone network for real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 752–760.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[30] Z. Rafii, A. Liutkus, F. R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-A corpus for music separation," 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[32] F. R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Proc. Latent Variable Anal. Signal Separation*, 2018, pp. 293–305.