



DiffStereo: End-to-End Mono-to-Stereo Audio Generation with Diffusion Transformer

Suqi Zhang¹, Zheqi Dai¹, Yongyi Zang², Yin Cao³, Qiuqiang Kong^{*1}

¹The Chinese University of Hong Kong, Hong Kong SAR, China

²Independent Researcher, Seattle, WA, USA

³Xi'an Jiaotong Liverpool University, Suzhou, China

1155226708@link.cuhk.edu.hk, zheqidai@cuhk.edu.hk, zyy0116@gmail.com,
yin.k.cao@gmail.com, qqkong@ee.cuhk.edu.hk

Abstract

Mono-to-stereo audio generation is a task that converts mono audio into two-channel stereo audio. Stereo audio generation plays a crucial role in enhancing spatial perception and auditory immersion. Traditional mono-to-stereo methods include rule-based, simulation-based, and various deep learning-based approaches. These methods require expert knowledge or explicit positional information to generate specific stereo effects, limiting their scalability and generalization. To address these challenges, we propose DiffStereo, an end-to-end diffusion transformer-based model that generates stereo audio conditioned on mono audio. The contributions of DiffStereo are as follows: First, DiffStereo directly synthesizes stereo audio from a mono waveform input in an end-to-end fashion, requiring no human intervention or prior knowledge. Second, DiffStereo achieves competitive objective ratings and consistently better subjective ratings, validating the effectiveness of our end-to-end approach.

Index Terms: Mono-to-stereo audio generation, spatial audio generation, diffusion transformer (DiT)

1. Introduction

Mono-to-stereo is the task of generating two-channel audio from a single channel. Stereo audio generation is essential for enhancing spatial perception and auditory immersion. By incorporating left and right channels, stereo audio produces sound that can enrich the listening experience. Stereo audio has diverse applications, including film viewing and music listening. When using headphones, stereo sound provides an immersive auditory environment that significantly enhances user engagement. Moreover, Stereo audio generation has applications to generate mono recordings into richer stereo representations. Stereo audio generation can also be used to remake history mono audio recordings. Consequently, the task of mono-to-stereo upmixing is becoming increasingly vital in achieving these objectives.

The evolution of mono-to-stereo upmixing in music information retrieval (MIR) developed rapidly over time. Traditional approaches [1] relied on complex signal processing techniques and require substantial human intervention and digital signal processing (DSP) expertise. Previous mono-to-stereo systems apply instrumental extraction methods to extract audio sources from mono recordings through deep learning approaches [2, 3, 4, 5] [6]. However, those methods suffer from audio artifacts. Recent advances in machine learning have adopted neural networks to automatically upmix music. However, those methods often rely on prior knowledge about the

stereo field. In Serrà et al.'s ISMIR study [7], parametric stereo techniques are used to reconstruct stereo sound. Recent Ambisonizer system [8] was proposed to generate first-order ambisonics signals. This reliance on pre-defined stereo space priors limits their overall performance potential. As end-to-end approaches show promise in many deep learning applications, modern generative models [9] has become instrumental in stereo/spatial audio generation [10, 11, 12, 13]. Recently, diffusion models [14] was proposed to address the audio denoising problem. However, these models typically rely on conditional information from various additional modalities, including text, melodic features, and spatial positions [15, 16, 17, 18]. However, obtaining such information is difficult in practice, which limits the availability of training data for these models.

In this work, we present DiffStereo¹, an end-to-end stereo audio generative model leveraging the diffusion transformer (DiT) architecture [19]. DiffStereo is trained on stereo audio only, where their mono counterparts are used to directly reconstruct stereo signals. Our approach makes no assumptions about the stereo audio space and requires no additional annotations or modalities to encode stereo information. Through empirical evaluation, we demonstrate that DiffStereo achieves comparable performance with both traditional and deep learning-based methods, while consistently achieving superior subjective ratings compared to prior approaches. To our knowledge, DiffStereo represents the first method capable of directly generating stereo recordings from mono signals. Given the established scalability of the DiT architecture, we believe our work provides a promising foundation for future research in mono-to-stereo conversion at scale.

This paper is structured as follows. In Section 2, we present our methodology, followed by a description of our experimental setup in Section 3. We report our findings in Section 4. We summarize our contributions and future directions in Section 5.

2. Methodology

In this study, we focus on end-to-end stereo audio generation, aiming to generate stereo audio from a given mono audio signal, thereby enhancing the mono audio with stereo effects to achieve binaural auditory perception. The proposed method adopts the diffusion transformer (DiT) directly in frequency domain to learn the distribution of stereo effects.

2.1. Complex Spectrum Input and Condition

Unlike previous diffusion model-based mono-to-stereo methods, such as Diff-SAGe [15] and ImmerseDiffusion [16], which require additional inputs such as text, images, or spatial lo-

*Corresponding author

¹<https://github.com/SAKI-77/DiffStereo>

cations, our model derives its condition by averaging the two stereo channels. Given the stereo audio $s \in \mathbb{R}^{2 \times N}$ with left and right channels s^l and s^r , and a time length N , we compute its corresponding mono representation as $m = (s^l + s^r)/2 \in \mathbb{R}^{1 \times N}$ and use it as the conditioning information for our model.

We apply short-time Fourier transform (STFT) to both the input s and the condition m , and then use a logarithmic transformation to pre-process the spectrograms and reduce the wide range of magnitudes.

$$S' = \gamma \cdot \log(1 + |S|) \cdot e^{j \cdot \angle S} \quad (1)$$

Where S is the complex spectrum of s and $\angle(\cdot)$ represents the phase. The logarithmic transformation compresses the dynamic range, making the data easier for neural networks to process and helping the model capture spectral features that align better with human auditory perception. γ is a non-negative real number factor that normalizes the amplitudes within the range $[0, 1]$. This ensures consistent scaling of the input data relative to the Gaussian diffusion noise [14], preventing extreme values from dominating the training. We then train our model on these pre-processed data to help the model learn more effectively from the spectral features.

2.2. Diffusion Transformer (DiT)

Diffusion transformers (DiTs) use the standard transformer framework as the backbone for diffusion models, achieving better scalability and performance than U-Net-based diffusion models. It comprises three core components:

2.2.1. Patchify Mechanism

The noised latent representation from a variational autoencoder (VAE), which is input, is divided into $p \times p$ patches and converted into a sequence of tokens. The number of tokens T depends on the patch size p , with smaller patches increasing both sequence length and computational cost. Positional embeddings (sine-cosine) are applied to all tokens.

2.2.2. Transformer Blocks and Conditioning

DiT processes input tokens through a series of transformer blocks. Conditional information, such as noise timesteps, class labels, or text prompts, can be incorporated using different strategies. Among these, an improved version of adaptive layer norm (adaLN) block called adaLN-Zero demonstrates superior performance.

This approach enhances adaptive layer normalization by incorporating zero-initialization to improve stability and training efficiency. Instead of fixed normalization parameters, it predicts the scale (γ) and shift (β) from the combined embedding of conditional inputs. It also introduces a learnable scaling factor (α), applied before residual connections within the transformer block. Inspired by techniques in ResNets and diffusion U-Nets, this initialization helps preserve identity mappings early in training, leading to better convergence and performance.

2.2.3. Transformer Decoder

After passing through the transformer blocks, a linear decoder reconstructs the noise prediction and covariance estimation. The output sequence is reshaped back into the original spatial format to complete the diffusion process.

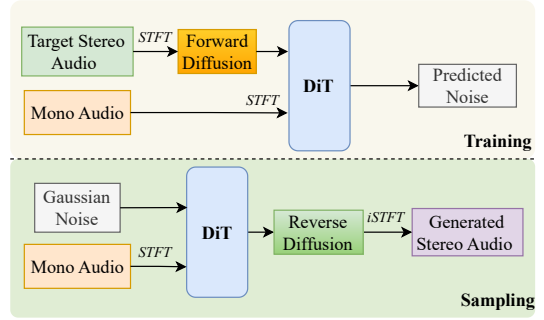


Figure 1: Overall Training/Sampling Pipeline of DiffStereo.

2.3. DiffStereo: Diffusion Transformer based Generative Model

Our model is built upon the transformer-based diffusion model, designed to operate directly on complex spectrograms. Fig.1 shows the overall training and sampling pipeline of the proposed model.

The diffusion model consists of a forward diffusion process where noise is incrementally introduced into the input data, and a reverse diffusion process in which the model learns to denoise and restore the original signal. Specifically, given that the input x_0 is stereo audio and the condition is its corresponding mono audio, in the forward stage, Gaussian noise is gradually added to the complex spectrum of stereo audio at each timestep following the formulation:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I), \quad (2)$$

where x_t denotes the input noisy complex spectrum at timestep t , and β_1, \dots, β_T denotes to a sequence of predetermined noise schedule parameters that control the noise levels. As noise accumulates over time, the complex spectrum x_t eventually becomes indistinguishable from pure Gaussian noise $\varepsilon \sim \mathcal{N}(0, I)$ at the final step T .

The learning objective of diffusion is to gradually remove these noises that added to the input and restore the original audio, which is called denoising process. During training, diffusion models learn this process with learnable parameters θ , which is defined as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (3)$$

The model is trained to learn this reverse process by minimizing the mean squared error (MSE) between the actual noise and the predicted noise, which is expressed as:

$$\mathcal{L}_{\text{Diff}} = \mathbb{E}_{x, \varepsilon \sim \mathcal{N}(0, I), t} [\|\varepsilon - \varepsilon_\theta(x_t, t, c)\|^2], \quad (4)$$

where c is the conditional mono audio. In sampling, the model estimates both the mean μ_θ and variance Σ_θ using the predicted noise to iteratively reconstruct the clean complex spectrum.

DiT serves as the foundational framework in our model, which adopts a transformer architecture with improvements such as stacked self-attention layers, gated MLPs, and skip connections. The self-attention mechanism is crucial for modeling complex audio structures, as it effectively captures long-range dependencies within the complex spectrum. Meanwhile, gated MLPs enhance the model's ability to learn complex non-linear

relationships, thereby improving its generative capabilities. Positional embeddings are utilized to aid the model in understanding the temporal order of complex spectrum frames.

3. Experimental Setup

3.1. Dataset

Our network is trained on the MUSDB18HQ [20] dataset, which consists of a total of 150 full-track songs of different styles and includes both the stereo mixtures and the original sources, divided between a training subset and a test subset. The train dataset composed of 100 songs, and the test set composed of 50 songs. We evaluate our approaches on the this test set.

3.2. Network Configurations

We use AdamW [21] with a weight decay of 0.001. The learning rate is set to 1×10^{-4} and audio length is 9.1 seconds. For experiment in frequency domain, we use Short-Time Fourier Transform (STFT) with hop size set to 300, n_{fft} set to 1024, and window length set to 1000. The scaling factor γ in Eq.1 is set to 0.15. We train our model for 96 hours on two NVIDIA 4090 GPUs with a batch size of 16 for around 80k epochs. The model is trained using complex spectrum of audio and the DiT-S-8 configuration, where the depth of transformer layer is set to 12, the number of attention head is set to 6, and the hidden size of the model is set to 384.

3.3. Evaluation Metrics

We apply both objective and subjective metrics to evaluate the mono-to-stereo audio generation performance. The objective metrics includes log-spectral distance [22], scale-invariant signal-to-noise ratio (SI-SNR) [23], wideness analysis, and Fréchet Audio Distance (FAD) [24] scores. The subjective metrics include mean opinion scores (MOS).

- **Log-Spectral Distance (LSD):** We compute the root-mean-square difference between two log-magnitude spectrograms, averaged across both frequency and time. We evaluate both *Mid/Side* LSD and *Left/Right* LSD metrics.
- **SI-SNR:** SI-SNR evaluates the quality of the predicted signal by comparing it with the clean target while ignoring differences in overall amplitude. It is calculated separately for the *Mid/Side* and *Left/Right* components. A higher SI-SNR suggests that more of the original signal has been preserved with less noise or distortion.
- **Wideness Analysis:** We evaluate 1) *Normalized Width* to indicate how wide the stereo field appears. A higher value suggests a wider perceived soundstage; 2) *Channel Correlation* to measure the similarity between the left and right channels. Lower correlation indicates a more spacious and immersive stereo effect; and *Phase Correlation* to capture the phase differences between channels.
- **FAD (Fréchet Audio Distance):** FAD compares the distribution of embeddings from generated audio with those from reference audio. It is calculated separately for the *Mid/Side* and *Left/Right* components to provide a more comprehensive evaluation. A lower FAD indicates better perceptual similarity to the ground-truth audio.
- **MOS:** In addition to these objective metrics, we conduct subjective evaluations of MOS, where human listeners rate the perceived stereo effects of the generated stereo audio on a scale, typically ranging from 1 (poor) to 5 (excellent).

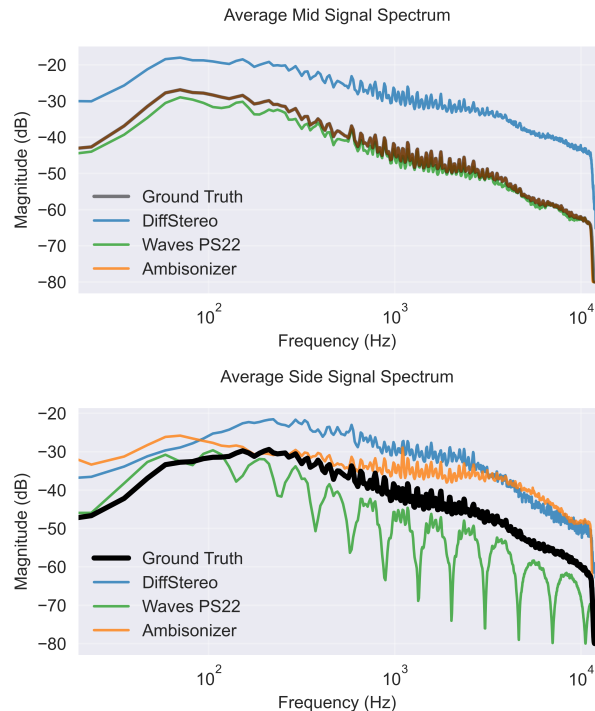


Figure 2: *Mid and Side Spectrum of signal.*

4. Results

4.1. Baseline Models

We compare our proposed DiffStereo framework against two baseline models. **Waves PS22 Stereo Maker**, a commercial VST plugin, converts mono to stereo audio through DSP techniques by introducing artificial phase perturbation to create the stereo field. **Ambisonizer** [8], trained on the MUSDB18HQ dataset with additional ambisonic impulse response augmentation, offers channel-agnostic neural upmixing based on spherical harmonics. It can generate first-order Ambisonic (FOA) audio from mono or stereo inputs. The model enforces ambisonic relationships during training, producing W, X, and Y ambisonic channels. Following the original paper’s approach, we convert these to stereo by performing a grid search at 10-degree intervals to find the position where left and right channels exhibit minimal RMS difference, accounting for phantom center bias in stereo audio.

4.2. Visualization of mono-to-stereo systems spectrum

The top part of Fig.2 shows the spectrum of mid signal computed by averaging left and right channel signals $Mid(t) = (L(t) + R(t))/2$. The bottom part of Fig.2 shows the spectrum of side signal computed as the difference between the left and right channel signals $Side(t) = (L(t) - R(t))/2$.

Fig.2 shows that the Waves PS22 model exhibits noticeable comb filtering artifacts in the side signal spectrum, likely due to its use of correlation-based processing. In contrast, the Ambisonizer model demonstrates competitive performance with a natural spectral shape; however, despite the loudness normalization at segment-level, its frame-to-frame loudness falls short compared to the ground truth, which might affect the perceived quality of the audio despite its natural spatial characteristics.

Table 1: Objective Evaluation Metrics for DiffStereo and Baseline Models.

Model	Log Spectral Distance ↓				SI-SNR (dB) ↑			
	Mid	Side	Left	Right	Mid	Side	Left	Right
Waves PS22	0.3951	0.7842	0.5442	0.5236	-22.94	-24.55	-20.87	-18.03
Ambisonizer	0.4351	0.6572	0.4550	0.4580	47.78	-31.57	4.64	4.93
DiffStereo	0.6324	0.7578	0.6806	0.6933	4.96	-44.97	1.07	0.96

Table 2: Wideness Analysis and FAD Metrics.

Model	Wideness Analysis			FAD ↓			
	Norm Width	Ch Corr	Phase Corr	Mid	Side	Left	Right
Ground Truth	0.715	0.761	0.238	-	-	-	-
Waves PS22	0.966	0.592	0.408	11.92	10.69	11.36	10.93
Ambisonizer	0.910	0.653	0.347	0.00	0.62	0.18	0.14
DiffStereo	0.867	0.676	0.324	0.20	1.53	0.33	0.33

The DiffStereo model, on the other hand, closely matches the spectral characteristics of the ground truth, particularly in both the mid and side signal spectra. Additionally, the higher frame-to-frame loudness level of DiffStereo may contribute to its preference in subjective evaluations [25], as listeners tend to favor louder audio when other quality aspects are comparable.

4.3. Objective Metrics

Table 1 shows the LSD and SI-SNR across both mid, side, left and right channels. For all but side channels, DiffStereo exhibits larger distance metric, which given that DiffStereo is a generative model, is expected and acceptable. When comparing the SI-SNR results, Ambisonizer demonstrates the highest quality among the generated stereo signals, with a mid-signal SI-SNR reaching 47.78 dB, indicating a strong preservation of the original audio characteristics. DiffStereo follows as the second-best performer, while Waves PS22 exhibits the poorest performance in this regard.

Table 2 shows the wideness analysis and FAD metrics, DiffStereo achieves a normalized width of 0.867, which is closer to the ground truth (0.715) compared to the other models. Additionally, it maintains a channel correlation of 0.676, suggesting more perceptually coherent stereo content. Ambisonizer achieves the lowest FAD scores, with 0.18 and 0.14 for the left and right channels, respectively. While both Waves PS22 and Ambisonizer leverage physical priors in their processing methodologies, their outputs closely align with ground-truth distributions. The Ambisonizer approach notably preserves mid-channel content intact, focusing exclusively on side-channel generation. However, despite Ambisonizer’s strong performance metrics, its practical applications may be limited by two factors: the requirement for grid search at inference time and tendency to produce an artificially wide stereo field. In contrast, DiffStereo, implementing an end-to-end generative architecture, demonstrates comparable Fréchet Audio Distance (FAD) to existing state-of-the-art methods while maintaining the ability to synthesize novel stereo content.

4.4. Subjective Metrics

We apply three MOS metrics, including the *naturalness*, the *overall quality*, and the degree of *perceived spatialization* of generated stereo audios as evaluation metrics. For the naturalness and quality metrics, participants provided discrete ratings

Table 3: MOS Test Results on Generated Stereo Audio.

Model	Naturalness	Quality	Spatial
Waves PS22	3.683 ± 1.147	3.65 ± 1.077	3.8
Ambisonizer	3.917 ± 0.759	3.833 ± 0.756	3.4
DiffStereo	3.45 ± 1.189	3.083 ± 1.173	4.6

on a 5-point scale, with higher scores indicating superior perceived quality. The study encompassed 12 participants, each of whom evaluated 5 randomly selected audio sets. Each evaluation set consisted of the original mono recording alongside its corresponding stereo versions generated by three different models.

Table 3 shows that DiffStereo significantly outperforms all baselines in spatial MOS (4.6), demonstrating its capability to synthesize the most immersive stereo experience. For naturalness MOS, Ambisonizer achieves the highest score, followed by Waves PS22, while DiffStereo shows slightly lower performance. This suggests a potential tradeoff between spatialization and naturalness, where stronger spatial effects may come at the expense of perceived naturalness. Overall, DiffStereo demonstrates strong spatialization capabilities while maintaining competitive MOS scores, even without leveraging additional priors, which further validates the effectiveness of our approach.

5. Conclusion

In this work, we proposed DiffStereo, an end-to-end diffusion transformers based model that could generate stereo audio conditioned on mono audio without extra types of condition information. Objective evaluations show that DiffStereo achieves a relatively high normalized width and competitive FAD scores, indicating its ability to synthesize spatially immersive and perceptually coherent stereo audio. The model also achieves high MOS ratings in terms of spatial effect, demonstrating its effectiveness in capturing realistic stereo imaging. However, our results also highlight a limitation: the model exhibits lower Quality MOS scores. To address this issue, future work will explore more complex configurations, such as DiT-L-8, and investigate the integration of diffusion transformers to improve performance and scalability.

6. References

- [1] M. Lagrange, L. G. Martins, and G. Tzanetakis, "Semi-automatic mono to stereo up-mixing using sound source formation," *Journal of The Audio Engineering Society*, 2007.
- [2] H. Liu, Q. Kong, and J. Liu, "Cws-presunet: Music source separation with channel-wise subband phase-aware resunet," *arXiv preprint arXiv:2112.04685*, 2021.
- [3] Y. Luo and J. Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.
- [4] W.-T. Lu, J.-C. Wang, Q. Kong, and Y.-N. Hung, "Music source separation with band-split rope transformer," in *Proc. ICASSP*, 2024, pp. 481–485.
- [5] S. Venkatesh, A. Benilov, P. Coleman, and F. Roskam, "Real-time low-latency music source separation using hybrid spectrogram-tasnet," in *Proc. ICASSP*, 2024, pp. 611–615.
- [6] D. Fitzgerald, "Upmixing from mono-a source separation approach," in *Proc. International Conference on Digital Signal Processing*, 2011, pp. 1–7.
- [7] J. Serrà, D. Scaini, S. Pascual, D. Arteaga, J. Pons, J. Breebaart, and G. Cengarle, "Mono-to-stereo through parametric stereo generation," *arXiv preprint arXiv:2306.14647*, 2023.
- [8] Y. Zang, Y. Wang, and M. Lee, "Ambisonizer: Neural up-mixing as spherical harmonics generation," *arXiv preprint arXiv:2405.13428*, 2024.
- [9] Z. Dai, H. He, and Q. Kong, "Musimple: A simplified music generation system with diffusion transformer," in *Proc. ICASSPW*, 2025, pp. 1–5.
- [10] Y. Chen, K. Shimada, C. Simon, Y. Ikemiya, T. Shibuya, and Y. Mitsufuji, "Cstereo: Audio-visual contextual and contrastive learning for binaural audio generation," *arXiv preprint arXiv:2501.02786*, 2025.
- [11] K. K. Parida, S. Srivastava, and G. Sharma, "Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3347–3356.
- [12] Z. Li, B. Zhao, and Y. Yuan, "Cross-modal generative model for visual-guided binaural stereo generation," *Knowledge-Based Systems*, vol. 296, p. 111814, 2024.
- [13] R. Dagli, S. Prakash, R. Wu, and H. Khosravani, "See-2-sound: Zero-shot spatial environment-to-spatial sound," *arXiv preprint arXiv:2406.06612*, 2024.
- [14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [15] S. Singh Kushwaha, J. Ma, M. R. Thomas, Y. Tian, and A. Bruni, "Diff-sage: End-to-end spatial audio generation using diffusion models," *arXiv e-prints*, pp. arXiv–2410, 2024.
- [16] M. Heydari, M. Souden, B. Conejo, and J. Atkins, "Immersedif-fusion: A generative spatial audio latent diffusion model," *arXiv preprint arXiv:2410.14945*, 2025.
- [17] P. Sun, S. Cheng, X. Li, Z. Ye, H. Liu, H. Zhang, W. Xue, and Y. Guo, "Both ears wide open: Towards language-driven spatial audio generation," *arXiv preprint arXiv:2410.10676*, 2024.
- [18] A. Levkovitch, J. Salazar, S. Mariooryad, R. Skerry-Ryan, N. Bar, B. Kleijn, and E. Nachmani, "Zero-shot mono-to-binaural speech synthesis," *arXiv preprint arXiv:2412.08356*, 2024.
- [19] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.
- [20] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimi-lakis, and R. M. Bittner, "MUSDB18-HQ - an uncompressed version of MUSDB18," 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:212982970>
- [21] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [22] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.
- [23] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR—half-baked or well done?" in *Proc. ICASSP*, 2019, pp. 626–630.
- [24] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2018.
- [25] E. Vickers, "The loudness war: Background, speculation, and recommendations," in *Proc. Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.