

PRESERVING SPATIAL INFORMATION IN MUSIC SOURCE SEPARATION

M. Acerbi*, M. Pezzoli, F. Antonacci

L. Bianchi

Politecnico di Milano, Dipartimento di Elettronica,
Informazione e Bioingegneria (DEIB)
Milano, Italy

Elettromedia S.p.A.,
Potenza Picena (MC), Italy

ABSTRACT

Music Source Separation (MSS) models based on Deep Neural Networks have reached state-of-the-art performance in splitting stereo mixtures into separate sources. However, little attention has been paid to how well these models preserve the spatial properties of each separated source. Motivated by applications in virtual/augmented reality and accessibility, we propose Spatially-Aware HT-Demucs (SA-HTDemucs), a novel deep learning framework that extends HT-Demucs with a per-source SpatialCueModule, which corrects the Interaural Level Difference (ILD) in the time-frequency domain to preserve the azimuth panning of each estimated source. The model is trained and evaluated on *binauralMUSDB18-HQ*, a novel binaural dataset synthesized by convolving each source stem in *MUSDB18-HQ* with open-source Head-Related Transfer Functions (HRTFs) at randomized horizontal positions. Results show that SA-HTDemucs consistently improves spatial location preservation across all sources compared to the HT-Demucs baseline, opening promising directions at the intersection of MSS and immersive audio.

Index Terms— Music Source Separation, Binaural Audio, Interaural Level Difference, Spatial Cue Preservation, Deep Neural Networks

1. INTRODUCTION

The growing demand for immersive audio experiences in virtual and augmented reality, music production, and accessibility has renewed interest in Music Source Separation (MSS), the task of decomposing a music mixture into its constituent sources, typically *vocals*, *drums*, *bass*, and *other* accompaniments [1]. This problem has attracted significant research interest, driven by applications in demixing, upmixing, and music information retrieval [2, 3, 4]. Deep Neural Networks (DNNs) proved their ability to model complex spectral and temporal patterns in audio signals, leading to substantial improvements in separation quality [5, 6]. Early landmark systems such as Open-Unmix [7] and Spleeter [8]

operated in the spectrogram domain, applying learned masks on the STFT magnitude, while waveform-domain architectures such as TasNet [9], Conv-TasNet [10] and Demucs [11] advanced the state of the art by processing raw audio signals directly. More recently, HT-Demucs [12] achieved state-of-the-art results on MUSDB18-HQ [13] by combining a hybrid time-frequency U-Net with a cross-domain Transformer.

Despite these advances, existing MSS models are typically evaluated solely on separation quality metrics such as SDR, without considering whether the spatial properties of each separated source are preserved. As shown in recent work [14], stereo MSS models largely fail to preserve the spatial information critical to binaural audio quality, with the degree of degradation depending on model architecture and target source. This gap is significant in binaural audio, where spatial perception is encoded through Interaural Level Difference (ILD), Interaural Time Differences (ITD), and HRTF filtering, and preserving such cues is essential for VR/AR, hearing aids, and accessibility [15, 16, 17]. Related work in speech has demonstrated this direction through binaural separation preserving interaural cues [18], joint separation and localization [19], and spatial loss functions [20].

Motivated by these results, this work proposes Spatially-Aware HT-Demucs (SA-HTDemucs), a novel DNN that extends HT-Demucs [12] with a per-source *SpatialCueModule*. Each module operates in the time-frequency domain and learns to correct the azimuth location of each estimated source based on ILD evaluation. The model is trained and evaluated on *binauralMUSDB18-HQ* [14], a novel binaural dataset synthesized by convolving each source stem in MUSDB18-HQ with open-source HRTFs at randomized azimuth angles on the horizontal plane, and summing the resulting binaural stems to form the mixture. SA-HTDemucs is compared against an HT-Demucs baseline, trained on the same dataset, using two complementary metrics: SI-SDR, to verify that the spatial correction stage does not degrade separation quality, and ILD MAE, an interaural cue-based metric proposed in this work to assess spatial preservation. Results show that SA-HTDemucs consistently reduces ILD errors across the entire audible frequency range for all the separated sources, with a moderate degradation in separation

*M. Acerbi is with Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB) and Elettromedia S.p.A.

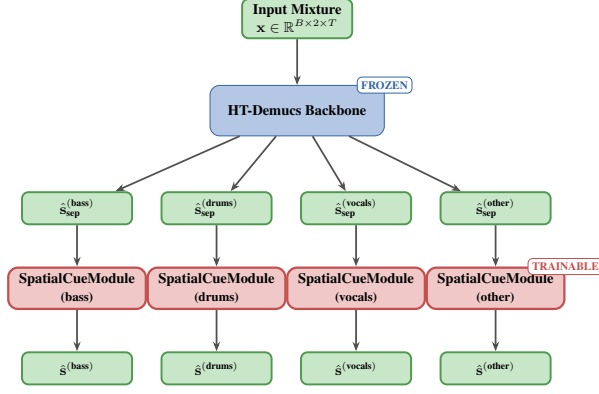


Fig. 1: SA-HTDemucs architecture.

quality with respect to the HT-Demucs baseline.

2. PROPOSED METHOD

Given a stereo binaural mixture $\mathbf{x} \in \mathbb{R}^{B \times 2 \times T}$, the MSS task estimates S source stems $\hat{\mathbf{s}}^{(s)} \in \mathbb{R}^{B \times 2 \times T}$, $s = 1, \dots, S$, such that $\mathbf{x} \approx \sum_{s=1}^S \hat{\mathbf{s}}^{(s)}$. In the binaural setting, an effective separation model should preserve the spatial location of each source as present in the original mixture. Among interaural cues, the ILD, defined as the per-sub-band energy ratio between left and right channels, is a well-established correlate of perceived azimuth [21] and constitutes the primary spatial feature targeted here. SA-HTDemucs decomposes the task into two stages:

$$\hat{\mathbf{s}}_{\text{sep}}^{(s)} = \text{HT-Demucs}(\mathbf{x}), \quad \hat{\mathbf{s}}^{(s)} = \mathcal{F}_{\theta_s}(\hat{\mathbf{s}}_{\text{sep}}^{(s)}), \quad (1)$$

where HT-Demucs [12] (frozen) produces $S=4$ preliminary source estimates, and $\mathcal{F}_{\theta_s}$ is a dedicated SpatialCueModule that corrects the ILD of each source independently.

2.1. SA-HTDemucs

SA-HTDemucs combines a pre-trained HT-Demucs backbone with a set of lightweight, trainable SpatialCueModules, one per output source (see Fig. 1). The HT-Demucs weights are kept entirely frozen throughout training, and only the SpatialCueModule parameters are updated. This makes the training data-efficient, as only spatially-annotated data is required to update the spatial correction stage, while separation quality is inherited from the pre-trained backbone by construction.

2.2. SpatialCueModule

Each separated source $\hat{\mathbf{s}}_{\text{sep}}^{(s)} \in \mathbb{R}^{B \times 2 \times T}$ is independently processed by a dedicated SpatialCueModule $\mathcal{F}_{\theta_s}$, a lightweight CNN operating entirely in the time-frequency domain (see Fig. 2).

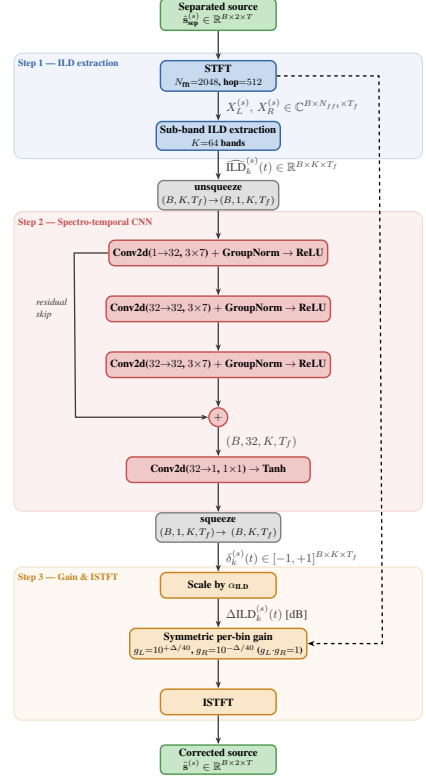


Fig. 2: SpatialCueModule architecture.

Step 1 – Sub-band ILD extraction.

A Short-Time Fourier Transform (STFT) is applied to the left and right channels of the separated source $\hat{\mathbf{s}}_{\text{sep}}^{(s)}$ ($N_{\text{fft}} = 2048$, $\text{hop} = 512$). The resulting spectra $X_L^{(s)}(f, t)$, $X_R^{(s)}(f, t) \in \mathbb{C}^{B \times N_{\text{fft}} \times T_f}$, where T_f is the number of time frames, are then partitioned into $K = 64$ mel-spaced frequency sub-bands. The per-band ILD trajectory is thus derived as

$$\text{ILD}_k^{(s)}(t) = 10 \log_{10} \frac{P_L^{(s)}(k, t)}{P_R^{(s)}(k, t)}, \quad (2)$$

where $P_c^{(s)}(k, t) = \sum_{f \in B_k} |X_c^{(s)}(f, t)|^2$ denotes the power of channel $c \in \{L, R\}$ in sub-band k at frame t .

Step 2 – Spectro-temporal CNN correction head.

The map $\text{ILD}_k^{(s)}(t) \in \mathbb{R}^{B \times K \times T_f}$ is reshaped to $\mathbb{R}^{B \times 1 \times K \times T_f}$ and processed by a three-layer spectro-temporal CNN with Conv2d kernels of size 3×7 (frequency $\times$ time), GroupNorm, and ReLU activations. A residual skip connection from the first to the third layer improves gradient flow, and the output projection is zero-initialized to ensure near-zero corrections at initialization. The resulting raw correction map $\delta_k^{(s)}(t) \in \mathbb{R}^{B \times K \times T_f}$ with values in $[-1, +1]$ is scaled by α_{ILD} to obtain

the ILD correction mask in dB as

$$\Delta\text{ILD}_k^{(s)}(t) = \alpha_{\text{ILD}} \cdot \delta_k^{(s)}(t). \quad (3)$$

Step 3 – Symmetric per-bin STFT gain application.

The correction is applied as a symmetric time-varying gain in the STFT domain. Since $\Delta\text{ILD}_k^{(s)}(t)$ is defined over $K=64$ sub-bands, each gain value is broadcast uniformly to all bins $f \in B_k$ of the full-resolution spectra $X_L^{(s)}(f, t)$, $X_R^{(s)}(f, t)$ as

$$\begin{aligned} g_L^{(s)}(f, t) &= 10^{\Delta\text{ILD}_k^{(s)}(t)/40}, \\ g_R^{(s)}(f, t) &= 10^{-\Delta\text{ILD}_k^{(s)}(t)/40}, \end{aligned} \quad (4)$$

where k is the sub-band index such that $f \in B_k$. By construction $g_L^{(s)}(f, t) \cdot g_R^{(s)}(f, t) = 1$ for all f and t , preserving the total loudness of $\hat{s}^{(s)}$. Finally, an ISTFT reconstructs the corrected waveform $\hat{s}^{(s)} \in \mathbb{R}^{B \times 2 \times T}$ left and right channels as

$$\begin{aligned} \hat{s}_L^{(s)} &= \text{STFT}^{-1} \left(g_L^{(s)}(f, t) \cdot X_L^{(s)}(f, t) \right), \\ \hat{s}_R^{(s)} &= \text{STFT}^{-1} \left(g_R^{(s)}(f, t) \cdot X_R^{(s)}(f, t) \right). \end{aligned} \quad (5)$$

2.3. Training Objective

Since the HT-Demucs backbone is frozen, only the $S=4$ SpatialCueModule weights are updated during training via a spatial supervision loss based on per-band ILD mean squared error (MSE) averaged over all S sources as

$$\mathcal{L} = \frac{1}{S} \sum_{s=1}^S \frac{1}{K \cdot T_f} \sum_{k=1}^K \sum_{t=1}^{T_f} \left(\widehat{\text{ILD}}_k^{(s)}(t) - \text{ILD}_{k, \text{gt}}^{(s)}(t) \right)^2, \quad (6)$$

where $\widehat{\text{ILD}}_k^{(s)}(t)$ is the per-band ILD of the corrected source $\hat{s}^{(s)}$ as output from the s -th SpatialCueModule, and $\text{ILD}_{k, \text{gt}}^{(s)}(t)$ is the per-band ILD of its ground truth counterpart $s_{\text{gt}}^{(s)}$. Both ILD are computed as in Eq. (2).

3. EXPERIMENTS AND RESULTS

The standard benchmark for MSS is MUSDB18-HQ [13], comprising 150 stereo tracks (100 training, 50 test), each with four isolated stems. To introduce explicit spatial information, we synthesize *binauralMUSDB18-HQ* following [14]. Each source stem is converted to mono and convolved with the left and right HRIRs from the SADIE II database [22] corresponding to a randomly selected azimuth angle $\theta \in [-90^\circ, +90^\circ]$ on the horizontal plane ($\phi = 0^\circ$), sampled in increments of 10° without replacement across sources. The resulting binaural stems are summed channel-wise and normalized to form the mixture. HRTF-based spatialization, rather than the amplitude panning common

in stereo music production, is consistent with the binaural, immersive audio applications targeted by SA-HTDemucs. Extending the evaluation to amplitude-panned material is left for future work. Of the 100 training tracks, 80 are used for training and 20 for validation; the 50-track test set is left unchanged. The baseline is HT-Demucs [12] fine-tuned on binauralMUSDB18-HQ for 500 epochs with L1 loss on the waveforms, without any explicit spatial supervision. To the best of our knowledge, no existing MSS method explicitly targets spatial cue preservation, making this the natural baseline to adopt: it isolates the contribution of the SpatialCueModule by showing what is achieved when the same backbone and training data are used without spatial supervision.

3.1. Implementation Details

SA-HTDemucs is trained with a batch size of $B=4$ on the binauralMUSDB18-HQ training set for 500 epochs using the Adam optimizer ($\text{lr}=10^{-3}$). The scaling factor $\alpha_{\text{ILD}}=15$ dB was determined empirically by analysing the distribution of the required ILD correction $|\text{ILD}_{k, \text{gt}}^{(s)}(t) - \widehat{\text{ILD}}_k^{(s)}(t)|$ across all training tracks: specifically, α_{ILD} was set to cover the 95th percentile of this distribution, ensuring that the SpatialCueModule output mask $\Delta\text{ILD}_k^{(s)}(t)$ is sufficient to represent the corrections required in the vast majority of cases. A systematic ablation over α_{ILD} is left for future work. The $S=4$ SpatialCueModules comprise approximately 700K trainable parameters, against the ~ 42 M frozen parameters of the HT-Demucs backbone. Training and inference were performed on a single NVIDIA TITAN RTX GPU. The source code and audio examples are publicly available online.¹

3.2. Evaluation Metrics

Results are evaluated on the 50-track test set using two evaluation metrics.

SI-SDR (Scale-Invariant Signal-to-Distortion Ratio).

Given the corrected source $\hat{s}^{(s)}$ and its ground truth $s_{\text{gt}}^{(s)}$, it measures the reconstruction quality of the separated waveform independently of global scale, as

$$\text{SI-SDR}(\hat{s}^{(s)}, s_{\text{gt}}^{(s)}) = 10 \cdot \log_{10} \frac{\left\| \frac{\langle \hat{s}^{(s)}, s_{\text{gt}}^{(s)} \rangle}{\|s_{\text{gt}}^{(s)}\|^2} \cdot s_{\text{gt}}^{(s)} \right\|^2}{\left\| \hat{s}^{(s)} - \frac{\langle \hat{s}^{(s)}, s_{\text{gt}}^{(s)} \rangle}{\|s_{\text{gt}}^{(s)}\|^2} \cdot s_{\text{gt}}^{(s)} \right\|^2}. \quad (7)$$

It is computed per source and then averaged over the test set to assess the separation quality impact of the spatial correction stage.

¹<https://github.com/polimi-ispl/sahtdemucs>

Table 1: SI-SDR (dB) comparison between SA-HTDemucs and the HT-Demucs baseline on the binauralMUSDB18-HQ test set (50 tracks).

Source	HT-Demucs (baseline)		SA-HTDemucs	
	Mean (dB)	Std (dB)	Mean (dB)	Std (dB)
Drums	+9.77	3.86	+ 9.30	3.67
Bass	+7.98	6.16	+ 7.58	5.98
Other	+4.30	3.83	+ 4.21	3.77
Vocals	+7.85	4.68	+ 7.56	4.56
All	+7.48	5.12	+ 7.16	4.95

ILD MAE (Mean Absolute Error on per-sub-band ILD).

The primary spatial metric is the frame-wise MAE between the per-sub-band ILD of the corrected source $\hat{s}^{(s)}$ and its ground truth counterpart $s_{\text{gt}}^{(s)}$, defined as

$$\text{MAE}_k(\hat{s}^{(s)}, s_{\text{gt}}^{(s)}) = \frac{1}{T_f} \sum_{t=1}^{T_f} \left| \widehat{\text{ILD}}_k^{(s)}(t) - \text{ILD}_{k, \text{gt}}^{(s)}(t) \right|, \quad (8)$$

where both ILD terms are computed as in Eq. (2).

Since each source stem in binauralMUSDB18-HQ is the result of the convolution with a fixed HRTF at a constant azimuth angle θ , the ground-truth ILD is stationary in time by construction. Therefore, the frame-wise average provides a robust estimate of the mean spatial position of each source in sub-band k , suppressing fluctuations due to musical content (e.g., note onsets, decays, and silent regions) that are unrelated to the source’s spatial position.

Table 1 reports the per-source SI-SDR comparison between SA-HTDemucs and the HT-Demucs baseline, evaluated for each of the four sources and averaged over the test set. The results reveal a moderate degradation in separation quality for SA-HTDemucs compared with the HT-Demucs baseline. This behavior is not unexpected: each SpatialCueModule applies a time-varying per-bin gain to the separated source in the STFT domain, which, while designed to correct the ILD, inevitably modifies the signal’s magnitude spectrum. Even though the gain is constrained to be symmetric ($g_L \cdot g_R = 1$) to preserve total loudness, the per-channel spectral modifications may introduce a degree of waveform distortion that penalizes SI-SDR.

Fig. 3 shows the per-band ILD MAE averaged over the test set for all four sources. The results demonstrate that SA-HTDemucs consistently improves spatial location preservation, reducing the ILD MAE across the entire audible frequency range relative to the HT-Demucs baseline. The per-band ILD MAE remains within 2–8 dB, whereas the baseline exhibits peaks of up to 16 dB, particularly in the 6–8 kHz region, which corresponds to the frequency range where head-shadow effects encoded in the HRTF are most prominent. This confirms that the SpatialCueModules effectively learn

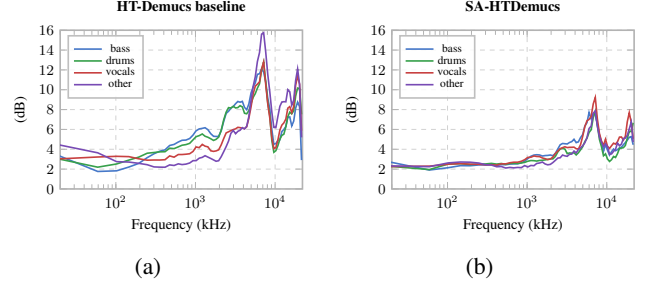


Fig. 3: ILD MAE on each source, averaged across binauralMUSDB18-HQ test set for (a) HT-Demucs baseline and (b) SA-HTDemucs.

to correct the azimuth panning of each separated source without requiring any modifications to the backbone architecture. Note that azimuth perception is dominated by ILD only above approximately 1.5 kHz, while at lower frequencies ITD becomes the primary cue [21]. The proposed ILD-only correction is thus most effective precisely where ILD is perceptually salient, consistent with the largest reductions observed above 5 kHz in Fig. 3. Low-frequency spatial fidelity thus rests entirely on the backbone and would require an ITD-aware, phase-sensitive counterpart.

4. CONCLUSIONS

In this work, we proposed SA-HTDemucs, a novel DNN that extends the HT-Demucs backbone with a per-source SpatialCueModule to preserve the azimuth panning of each separated source by correcting the per-sub-band ILD in the time-frequency domain. Unlike end-to-end approaches [19, 20], it casts spatial cue preservation as a post-processing correction on a frozen, pre-trained separator. Trained and evaluated on *binauralMUSDB18-HQ*, synthesized by convolving each source stem in MUSDB18-HQ with open-source HRTFs at randomized horizontal positions, SA-HTDemucs consistently reduces ILD MAE across the entire audible frequency range for all four sources, lowering peaks from 16 dB to within 2–8 dB. The moderate SI-SDR degradation with respect to the HT-Demucs baseline reveals an inherent trade-off between spatial fidelity and separation quality, which could be addressed through joint SI-SDR and ILD-based loss optimization. Further directions include complementing the ILD head with a phase-sensitive ITD correction (Sec. 3), dynamic panning scenarios, and integration into object-based upmixing pipelines for stereo-to-immersive rendering. Since the model is trained and evaluated primarily on ILD, a perceptual assessment and the extension to unseen HRTFs, real recordings, and further spatial cues (e.g., interaural coherence) remain important future work.

5. REFERENCES

- [1] Estefania Cano, Derry FitzGerald, Antoine Liutkus, Mark D Plumbley, and Fabian-Robert Stöter, “Musical source separation: An introduction,” *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 31–40, 2018.
- [2] Erkki Oja, Juha Karhunen, Harri Valpola, Jaakko Särelä, Mika Inki, Antti Honkela, Alexander Ilin, Karthikesh Raju, Tapani Ristaniemi, and Ella Bingham, “Independent component analysis and blind source separation,” *Helsinki Univ. Technol., Espoo, Finland, Tech. Rep.*, 2003.
- [3] Dan Barry, Robert Lawlor, and Eugene Coyle, “Real-time sound source separation: Azimuth discrimination and resynthesis,” 2004.
- [4] Agnieszka Roginska and Paul Geluso, *Immersive sound*, Focal Press, 2017.
- [5] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent, “Multichannel audio source separation with deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.
- [6] Daniel Turner and Damian T Murphy, “A deep learning approach to the prediction of time-frequency spatial parameters for use in stereo upmixing,” in *27th International Conference on Digital Audio Effects, DAFx 2024*. DAFx, 2024, pp. 428–435.
- [7] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji, “Open-unmix-a reference implementation for music source separation,” *Journal of Open Source Software*, vol. 4, no. 41, pp. 1667, 2019.
- [8] Romain Hennequin, Anis Khelif, Felix Voituret, and Manuel Moussallam, “Spleeter: a fast and efficient music source separation tool with pre-trained models,” *Journal of Open Source Software*, vol. 5, no. 50, pp. 2154, 2020.
- [9] Yi Luo and Nima Mesgarani, “Tasnet: time-domain audio separation network for real-time, single-channel speech separation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.
- [10] Yi Luo and Nima Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [11] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach, “Music source separation in the wave-form domain,” *arXiv preprint arXiv:1911.13254*, 2019.
- [12] Simon Rouard, Francisco Massa, and Alexandre Défossez, “Hybrid transformers for music source separation,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [13] Zafar Raffi, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, “Musdb18-hq - an uncompressed version of musdb18,” 2019.
- [14] Richa Namballa, Agnieszka Roginska, and Magdalena Fuentes, “Do music source separation models preserve spatial information in binaural audio?,” *arXiv preprint arXiv:2507.00155*, 2025.
- [15] Harald Viste and Gianpaolo Evangelista, “On the use of spatial cues to improve binaural source separation,” in *Proceedings of 6th International Conference on Digital Audio Effects (DAFx-03)*, 2003, pp. 209–213.
- [16] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang, “Self-supervised generation of spatial audio for 360 video,” *Advances in neural information processing systems*, vol. 31, 2018.
- [17] Thomas Potter, Z. Cvetković, and Enzo De Sena, “On the relative importance of visual and spatial audio rendering on vr immersion,” vol. 2, 2022.
- [18] Cong Han, Yi Luo, and Nima Mesgarani, “Real-time binaural speech separation with preserved spatial cues,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6404–6408.
- [19] Yu Chen, Bowen Liu, Zijian Zhang, and Hun-Seok Kim, “An end-to-end deep learning framework for multiple audio source separation and localization,” in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 736–740.
- [20] Wootae Lim and Juhan Nam, “Enhancing spatial audio generation with source separation and channel panning loss,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8321–8325.
- [21] Jens Blauert, *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997.
- [22] Cal Armstrong, Lewis Thresh, Damian Murphy, and Gavin Kearney, “A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database,” *Applied Sciences*, vol. 8, no. 11, pp. 2029, 2018.