Audio Engineering Society

Conference Express Paper

Presented at the AES 2026 International Conference on
Audio for Virtual and Augmented Reality and Immersive Games
2026 June 30 – July 3, Paris, France

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Complex Ratio Mask Ambisonics-to-Binaural Rendering with Intensity Vector Features and Perceptual Multi-Objective Optimization

Szymon Zaporowski[1] and Bartłomiej Mróz[1]

[1]*Department of Multimedia Systems, Gdańsk University of Technology*

Correspondence should be addressed to Szymon Zaporowski (szyzapor@pg.edu.pl)

ABSTRACT

This paper presents a neural network for binaural rendering of first-order Ambisonics (FOA) signals, enabling immersive spatial audio over headphones without individualized HRTF measurements at inference time. The model operates in the STFT domain using Complex Ratio Masks (CRM): a shared mask pair (left/right ear) is applied via complex-valued multiplication to all four FOA channels with directional weighting, preserving both magnitude and phase. The input representation extends standard spectral features with acoustic intensity vector channels encoding sound arrival direction at each time-frequency bin. Training uses a multi-objective loss combining SI-SDR, multi-resolution spectral reconstruction, and interaural level/phase difference terms. The four-level UNet backbone with channel-spatial attention totals approximately four million parameters. Evaluation against eight baseline configurations—spanning magnitude-masking variants (28 M parameters), a waveform-domain Wave-U-Net (143 M), and the A2B time-domain renderer—shows that the CRM variant achieves the strongest spectral and spatial fidelity in the cross-domain setting with a seven-fold parameter reduction over the magnitude-masking baseline while maintaining real-time CPU operation.

1 Introduction

Binaural rendering synthesizes a two-channel headphone signal that recreates the spatial impression of a three-dimensional sound scene. As VR, AR, and immersive gaming applications proliferate, demand for real-time binaural audio over consumer headphones has increased. First-order Ambisonics (FOA) is widely adopted for capturing and transmitting full-sphere soundfields due to its rotation-invariant encoding and compatibility with head-tracked playback [1].

The conventional FOA-to-binaural pipeline decodes the four spherical harmonic channels (W, Y, Z, X) onto virtual loudspeaker positions and convolves each signal with the corresponding Head-Related Transfer Function (HRTF) [2]. This approach suffers from two limitations: at first order, significant spatial aliasing arises above approximately 700 Hz, degrading localization accuracy; and perceptually accurate rendering requires individualized HRTF measurements that are impractical for mass-market deployment.

Data-driven approaches learn the FOA-to-binaural mapping from paired recordings without per-individual HRTF measurements. Zhu et al. [3] proposed an STFT-domain UNet predicting magnitude ratio masks on the omnidirectional channel W, but this approach discards phase information. Gebru et al. [4] introduced A2B, a time-domain WaveNet retaining full phase fidelity but imposing sequential inference cost. Neural methods have also been extended to Ambisonic encoding [5] and super-resolution [6].

Two gaps persist: (a) magnitude-mask methods discard inter-channel phase, while time-domain methods must implicitly learn spectral structure; (b) no architecture uses explicit directional FOA information as input. We address these gaps through three contributions: (1) a shared CRM pair applied via complex multiplication to all four FOA channels with directional weighting; (2) three intensity vector input channels providing per-bin sound arrival direction, yielding a 23-channel tensor; (3) a multi-objective perceptual loss combining SI-SDR, multi-resolution STFT reconstruction with perceptual frequency weighting, ILD/IPD terms, and time-domain L_1 distance.

The encoder-decoder backbone is a four-level UNet with channel-spatial attention ($\sim$4 M parameters). Evaluation on three test sets shows the CRM architecture achieves the strongest spectral and spatial scores in the cross-domain setting with a seven-fold parameter reduction and real-time CPU operation. Section 2 reviews related work, Section 3 describes the method, Section 4 the data, Section 5 the evaluation, and Section 6 concludes.

2 Related Work

2.1 Conventional Rendering

The standard FOA-to-binaural pipeline decodes the spherical harmonic channels onto virtual loudspeaker directions and convolves each with the corresponding HRTF [1]. Spatial fidelity is bounded by the Ambisonic order: at first order only four basis functions are available, causing spatial aliasing above $\sim$700 Hz. Individualized HRTF measurement requires specialized equipment; generic HRTF sets introduce coloration and localization biases that vary across listeners.

2.2 Neural STFT-Domain Approaches

Zhu et al. [3] proposed a UNet predicting real-valued magnitude ratio masks on the omnidirectional channel W, demonstrating the viability of neural spectral masking. However, the predicted masks modulate only magnitude, and phase is taken from the input W, preventing synthesis of fine interaural time differences critical for lateral localization below 1.5 kHz, where ITD cues dominate. The present work addresses this by predicting Complex Ratio Masks applied across all four FOA channels.

2.3 Neural Time-Domain Approaches

Time-domain renderers retain full phase information by operating on raw waveforms. Gebru et al. [4] introduced A2B, a WaveNet-based model ($<$700 k parameters) with competitive signal accuracy, but the autoregressive structure imposes sequential inference cost scaling linearly with output length. The proposed CRM architecture combines advantages of both paradigms by operating in the complex STFT domain, retaining frequency-domain efficiency while preserving phase.

2.4 Spatial Features and Loss Functions

The active acoustic intensity vector provides a per-bin directional estimate from FOA cross-spectra [7] but has not been used as a neural network input for binaural rendering. Multi-resolution STFT losses have proven effective in speech synthesis [8] and source separation [9] but have not been applied to this task. The present work combines intensity vector features with a multi-objective loss including multi-resolution STFT, ILD, and IPD terms.

3 Proposed Method

The system receives a four-channel FOA signal in ACN/SN3D ordering (W, Y, Z, X) and produces binaural output (L, R) in the STFT domain via Complex Ratio Masks applied to all four input channels.

3.1 CRM-Based Rendering Formulation

The network predicts a shared Complex Ratio Mask pair $\mathbf{M}_L, \mathbf{M}_R \in \mathbb{C}^{T\times F}$, each consisting of real and imaginary components, yielding a 16-channel output (4 FOA $\times$ 2 ears $\times$ 2 components). The output activation is $\tanh(\cdot)\times 2$, bounding mask values to $[-2,2]$. The binaural STFT for each ear $e \in \{L,R\}$ is:

$$\hat{Y}_e(t,f) = \mathbf{M}_e \odot W + \alpha \sum_{c\in\{Y,Z,X\}} \mathbf{M}_e \odot c, \qquad (1)$$

where $\odot$ denotes element-wise complex multiplication, W and c are the STFT representations of the respective FOA channels, W contributes at unit weight, and α controls the directional channel contribution (set to 1.0 in all experiments). Complex multiplication jointly modifies magnitude and phase, making this formulation more expressive than magnitude masking [10]. Although $\alpha = 1.0$ assigns equal weight to all channels in the current experiments, retaining α as an explicit hyperparameter separates the omnidirectional and directional contributions structurally, facilitating future tuning. The time-domain signal is recovered via inverse STFT.

3.2 Input Feature Representation

The input is a 23-channel tensor $[B, 23, T, F]$ computed from the FOA STFT ($n_{\text{fft}} = 1024$, hop $= 512$, Hann window): (1) log-magnitude spectrograms (4 ch), (2) cosine of phase (4 ch), (3) inter-channel phase differences for all six ordered channel pairs (C_i, C_j), each encoded as $|C_i|\cos(\angle C_i - \angle C_j)$ and $|C_i|\sin(\angle C_i - \angle C_j)$, where C_i is the first channel in the pair (12 ch). Weighting by $|C_i|$ means these features carry both phase-difference and magnitude information. Bins where the reference channel is loud contribute more, suppressing estimates in noise-dominated regions. The final group is (4) active intensity vector components $I_d(t,f) = \text{Re}\{W^*(t,f)\cdot D(t,f)\}$ for $D \in \{X,Y,Z\}$ (3 ch), providing an explicit per-bin estimate of sound arrival direction [7].

3.3 Multi-Objective Training Loss

Training minimizes a composite loss:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{SI-SDR}} + \lambda_2\mathcal{L}_{\text{STFT}} + \lambda_3\mathcal{L}_{\text{ILD}} + \lambda_4\mathcal{L}_{\text{IPD}} + \lambda_5\mathcal{L}_{\text{L1}}, \qquad (2)$$

with $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 0.15$, $\lambda_4 = 0.15$, $\lambda_5 = 0.2$, selected via preliminary validation experiments. The SI-SDR term [11] (negated) measures waveform-level accuracy. The multi-resolution STFT loss computes frequency-weighted log-magnitude spectral distance at three FFT sizes (256, 1024, 4096 samples): the short window captures transients, the medium targets speech/mid-frequency detail, and the long enforces low-frequency envelope fidelity, each emphasizing the 1–8 kHz range. The ILD loss computes interaural level difference error across five sub-bands (20 Hz to Nyquist); the IPD loss penalizes interaural phase difference error below 1.5 kHz, where phase-based localization cues are perceptually dominant; and an L1 term provides sample-level reconstruction.

3.4 Network Architecture and Training

The backbone is a four-level 2-D convolutional UNet (encoder widths 32, 64, 128, 256). Each level has a residual block (two 3×3 convolutions, batch normalization, LeakyReLU 0.01, identity/1×1 projection), CBAM-style channel-spatial attention [12] (reduction ratio 8, 7×7 spatial kernel), and average pooling (stride 2). The bottleneck applies one residual block at 256 channels with dropout ($p = 0.1$). The decoder uses transposed-convolution upsampling with skip connections. The output head: Conv2d$(32\to16, 3\times3)$ then Conv2d$(16\to16, 1\times1)$ followed by the $\tanh(\cdot)\times2$ activation described in Section 3.1; total $\sim$3.9 M parameters.

Training uses Adam (lr $= 3\times10^{-4}$, weight decay 10^{-5}), cosine annealing warm restarts ($T_0 = 20$, $T_{\text{mult}} = 2$), gradient clipping at 5.0, batch size 32, up to 200 epochs with early stopping (patience 30). Augmentation: inter-sample mixup (max strength 0.3) and additive Gaussian noise (probability 0.3, amplitude 0.005).

4 Dataset and Preprocessing

The training corpus pools four datasets of paired FOA–binaural recordings in ACN/SN3D format: the Zhu dataset [3] (31 train / 18 test files), the Argentum HOA corpus [13] (1,555 files, used exclusively for training), Urban Soundscapes of the World [14] (114 train / 19 test), and A2B 2-MP [4] (133 train / 23 test). The datasets span speech, urban soundscapes, and music, providing diverse acoustic content.

Table 1: Training corpus summary (5-second segments at 44.1 kHz).

Dataset	Train	Test	Segments
Zhu [3]	31	18	∼400
Argentum [13]	1,555	–	∼9,300
Urban Sound. [14]	114	19	∼1,400
A2B 2-MP [4]	133	23	∼3,900
Total	**1,833**	**60**	**∼15,000**

All audio files are loaded as 32-bit floating-point waveforms and resampled on-the-fly to 44.1 kHz using `torchaudio.functional.resample` when the native rate differs. Each recording is segmented into non-overlapping 5-second windows (220,500 samples); partial trailing segments are discarded. The total corpus comprises approximately 15,000 segments, split 80/20 for training/validation (Table 1). No input normalization is applied; amplitude range is implicitly constrained by the bounded tanh output activation. The STFT feature extraction ($n_{\text{fft}} = 1024$, hop $= 512$, Hann window) is performed within the model forward pass.

5 Evaluation

5.1 Systems and Protocol

Nine systems are compared across three families. The **Proposed CRM** is the architecture described in Section 3 (3.89 M parameters). A **Wave-U-Net** baseline adapts the one-dimensional fully convolutional architecture of Stoller et al. [15] to operate on the four-channel FOA waveform (142.61 M parameters); it is trained with the same SI-SDR, multi-resolution STFT, and L1 losses as the Proposed CRM.[1] The **Zhu [3] family** includes our re-implementation of the original architecture (27.64 M) plus five training variants: *Mixup* (mixup augmentation), *Spatial loss* (additional ILD/IPD terms), *Attention* (channel-spatial attention blocks, 27.95 M), *Attn (low-mix)* (attention with reduced mixup), and *PG data* (original architecture trained on a different corpus). As Zhu et al. did not release source code or pre-trained weights, all Zhu-family models are our own re-implementation.[2] The **A2B** [4] is the official time-domain WaveNet renderer (0.69 M, 48 kHz); for each test set we use the released checkpoint trained on the matching corpus where available.

Three test sets are used: A2B 2-MP (138 segments), UrbanSounds (114), and Zhu ByteDance (108), all 5-second windows. Metrics: SI-SDR, log-spectral distance (LSD), ILD error (5 sub-bands), IPD error ($<$1.5 kHz), PESQ, STOI, and BINAQUAL localization similarity [16]. Significance: two-sided Wilcoxon signed-rank at significance level 0.01.

5.2 Results

Tables 2–4 report the results; Figure 1 summarizes all metrics and Figure 2 shows SI-SDR distributions with confidence intervals.

5.3 Discussion

Matched checkpoints dominate. On the Zhu and UrbanSounds test sets, A2B with the matching checkpoint leads on every metric by a statistically significant margin ($p < 0.01$, Wilcoxon). On the Zhu set it reaches 10.49 dB SI-SDR, LSD of 7.41 dB, PESQ of 3.33, STOI of 0.87, and the highest BINAQUAL (0.19).

Cross-domain performance. On A2B 2-MP, where no matched A2B or Zhu checkpoint exists, both external baselines fail (A2B: -19.42 dB SI-SDR; Zhu original: -19.84 dB). The Proposed CRM and Wave-U-Net, trained on material including this corpus, lead with 4.89 dB and 5.29 dB respectively ($p < 10^{-3}$). However, the Proposed CRM retains the best spectral and spatial scores: lowest LSD (9.75 dB), lowest ILD error (3.50 dB), lowest IPD error (1.20 rad), and highest STOI (0.76), all significantly better than the Wave-U-Net despite its marginal SI-SDR advantage.

Training variants. The Zhu training variants do not transfer uniformly. On the Zhu test set the original checkpoint (6.52 dB SI-SDR, BINAQUAL 0.13) is far ahead of every variant, which collapse to between -23 and -40 dB; these variants were trained on a different corpus and do not generalize to the ByteDance

[1] Source code for the Proposed CRM and Wave-U-Net: `https://git.pg.edu.pl/p768047/avarig_2026_crm_ambisonics2binaural`

[2] Source code for the Zhu re-implementation: `https://git.pg.edu.pl/p768047/avarig_2026_zhu_implementation`

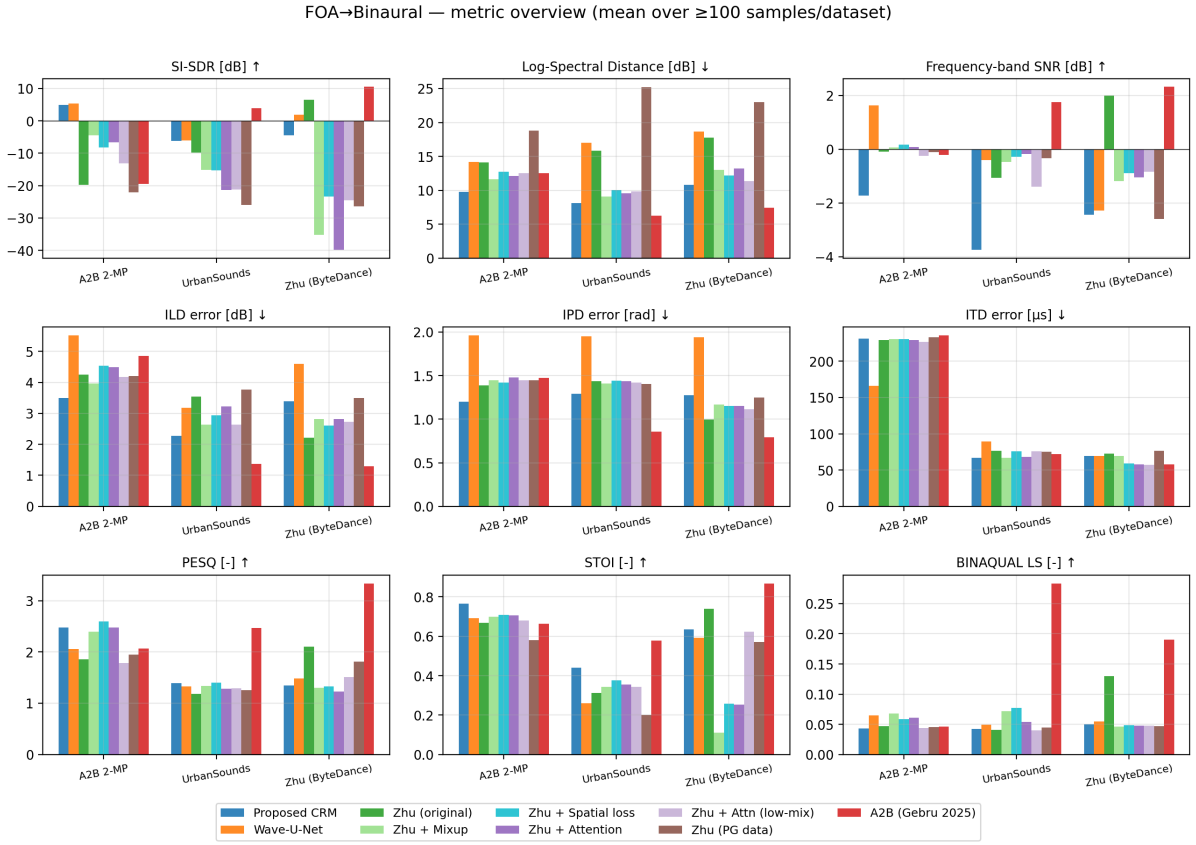


Fig. 1: Metric overview across three test sets (mean over $\geq$100 segments per set). Arrows indicate preferred direction.

Table 2: Results on the A2B 2-MP test set ($n = 138$). Mean $\pm$ 95% CI. Best in bold.

Model	SI-SDR ↑	LSD ↓	ILD err ↓	IPD err ↓	PESQ ↑	STOI ↑	BINAQUAL ↑
Proposed CRM	4.89±0.43	**9.75±0.29**	**3.50±0.15**	**1.20±0.04**	2.48±0.07	**0.76±0.02**	0.04±0.00
Wave-U-Net	**5.29±0.43**	14.18±0.55	5.52±0.22	1.96±0.03	2.06±0.06	0.69±0.02	0.06±0.00
Zhu (original)	-19.84±1.40	14.14±0.53	4.25±0.22	1.39±0.03	1.86±0.05	0.67±0.02	0.05±0.00
Zhu + Mixup	-4.47±0.66	11.62±0.40	3.97±0.18	1.45±0.03	2.40±0.06	0.70±0.03	**0.07±0.00**
Zhu + Spatial loss	-8.19±0.99	12.76±0.45	4.54±0.21	1.42±0.03	**2.59±0.06**	0.71±0.03	0.06±0.00
Zhu + Attention	-6.68±0.74	12.14±0.42	4.49±0.23	1.48±0.03	2.48±0.06	0.71±0.03	0.06±0.00
Zhu + Attn (l-mix)	-13.08±1.09	12.52±0.43	4.18±0.20	1.45±0.04	1.78±0.05	0.68±0.03	0.04±0.00
Zhu (PG data)	-22.01±1.45	18.80±1.04	4.20±0.23	1.45±0.03	1.95±0.12	0.58±0.03	0.05±0.00
A2B (Gebru 2025)	-19.42±1.16	12.50±0.41	4.86±0.20	1.47±0.02	2.07±0.06	0.66±0.02	0.05±0.00

test material. On A2B 2-MP, however, the mixup variant is the best member of the Zhu family (−4.47 dB) and is significantly better than the original (−19.84 dB, $p < 10^{-23}$), indicating that mixup augmentation improves cross-domain transfer even when overall accuracy remains limited.

Computational cost. Table 5 summarizes model size and inference speed. Systems span two orders of magnitude in size: from 0.69 M (A2B) to 142.61 M (Wave-U-Net). The Proposed CRM (3.89 M, RTF = 0.054) and Zhu family (RTF 0.04–0.06) are fastest; A2B is the only system exceeding real time on CPU (RTF≈1.37).

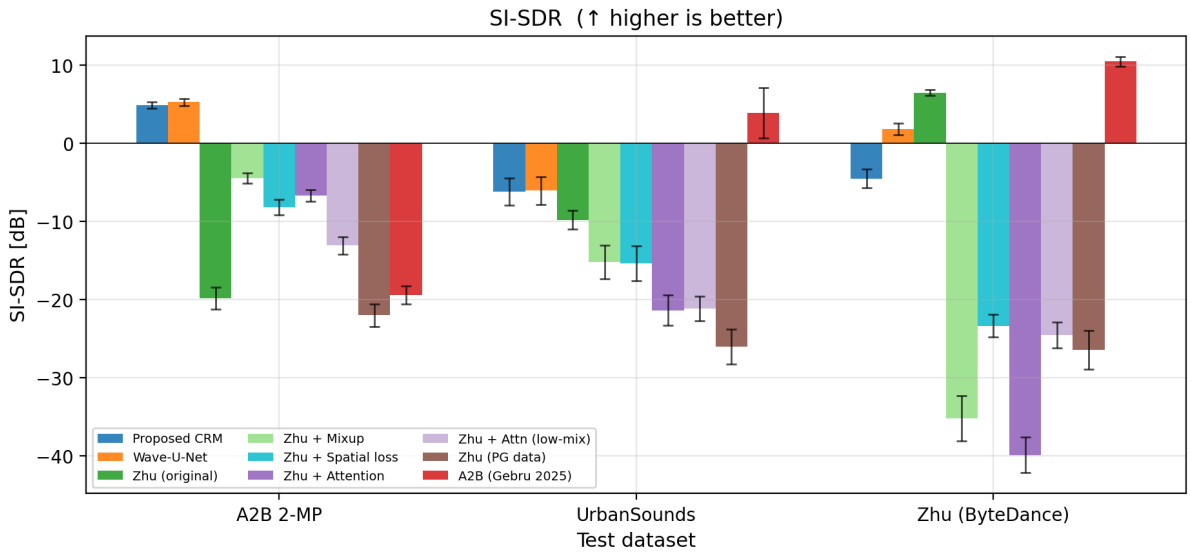


Fig. 2: SI-SDR with 95% confidence intervals. A2B with a matched checkpoint leads on the Zhu and UrbanSounds sets, whereas the Proposed CRM and Wave-U-Net lead on A2B 2-MP, where no matched A2B or Zhu checkpoint is available.

Table 3: Results on the UrbanSounds test set ($n = 114$). Mean ± 95% CI. Best in bold.

Model	SI-SDR ↑	LSD ↓	ILD err ↓	IPD err ↓	PESQ ↑	STOI ↑	BINAQUAL ↑
Proposed CRM	-6.16±1.75	8.13±0.24	2.28±0.18	1.29±0.02	1.39±0.09	0.44±0.04	0.04±0.00
Wave-U-Net	-6.04±1.79	17.00±0.72	3.18±0.19	1.95±0.03	1.33±0.07	0.26±0.03	0.05±0.00
Zhu (original)	-9.78±1.23	15.81±0.34	3.55±0.21	1.44±0.02	1.18±0.06	0.31±0.03	0.04±0.00
Zhu + Mixup	-15.19±2.16	9.06±0.32	2.64±0.14	1.41±0.02	1.33±0.12	0.34±0.03	0.07±0.01
Zhu + Spatial loss	-15.34±2.24	10.02±0.21	2.94±0.15	1.44±0.03	1.40±0.09	0.38±0.03	0.08±0.01
Zhu + Attention	-21.37±1.94	9.53±0.24	3.22±0.15	1.43±0.03	1.28±0.05	0.36±0.03	0.05±0.00
Zhu + Attn (l-mix)	-21.17±1.56	9.86±0.25	2.64±0.15	1.42±0.02	1.29±0.06	0.34±0.03	0.04±0.00
Zhu (PG data)	-26.03±2.22	25.19±1.80	3.77±0.25	1.40±0.02	1.25±0.08	0.20±0.03	0.04±0.00
A2B (Gebru 2025)	**3.90±3.21**	**6.25±0.51**	**1.37±0.17**	**0.86±0.10**	**2.47±0.38**	**0.58±0.06**	**0.28±0.06**

Table 4: Results on the Zhu (ByteDance) test set ($n = 108$). Mean ± 95% CI. Best in bold.

Model	SI-SDR ↑	LSD ↓	ILD err ↓	IPD err ↓	PESQ ↑	STOI ↑	BINAQUAL ↑
Proposed CRM	-4.49±1.17	10.80±0.45	3.39±0.13	1.27±0.02	1.35±0.04	0.64±0.02	0.05±0.00
Wave-U-Net	1.83±0.72	18.69±0.74	4.60±0.14	1.94±0.01	1.48±0.03	0.59±0.02	0.06±0.00
Zhu (original)	6.52±0.36	17.76±0.93	2.21±0.07	1.00±0.02	2.10±0.07	0.74±0.01	0.13±0.00
Zhu + Mixup	-35.18±2.89	13.02±0.73	2.81±0.09	1.17±0.02	1.30±0.07	0.11±0.06	0.05±0.00
Zhu + Spatial loss	-23.37±1.45	12.18±0.54	2.61±0.07	1.15±0.02	1.32±0.08	0.26±0.05	0.05±0.00
Zhu + Attention	-39.90±2.28	13.20±0.69	2.82±0.07	1.15±0.02	1.22±0.05	0.25±0.04	0.05±0.00
Zhu + Attn (l-mix)	-24.55±1.64	11.34±0.46	2.73±0.07	1.11±0.02	1.51±0.04	0.62±0.02	0.05±0.00
Zhu (PG data)	-26.45±2.49	22.98±2.02	3.49±0.21	1.25±0.02	1.81±0.09	0.57±0.04	0.05±0.00
A2B (Gebru 2025)	**10.49±0.60**	**7.41±0.19**	**1.29±0.04**	**0.79±0.02**	**3.33±0.09**	**0.87±0.01**	**0.19±0.00**

Table 5: Computational cost. RTF is mean CPU time / signal duration; RTF < 1 is faster than real time.

Model	Params [M] ↓	RTF ↓
Proposed CRM	3.89	0.054
Wave-U-Net	142.61	0.393
Zhu (original)	27.64	**0.038**
Zhu + Mixup	27.64	0.046
Zhu + Spat. loss	27.64	0.047
Zhu + Attention	27.95	0.054
Zhu + Attn (l-mix)	27.95	0.055
Zhu (PG data)	27.64	0.046
A2B (Gebru 2025)	**0.69**	1.370

The CRM offers the best efficiency–accuracy balance: a seven-fold reduction over Zhu at comparable speed with superior cross-domain spectral/spatial scores.

Perceptual localization. The BINAQUAL localization similarity corroborates the objective ranking where a matched checkpoint exists but also reveals that high SI-SDR does not guarantee faithful localization cues: on A2B 2-MP all systems score at or below 0.07 despite SI-SDR values near 5 dB for the leading models. Perceptual evaluation therefore adds information that signal-level and spectral metrics do not capture.

6 Conclusion

This paper presented a frequency-domain neural architecture for FOA-to-binaural rendering using shared Complex Ratio Masks applied to all four Ambisonics channels with directional weighting, augmented with acoustic intensity vector input features, and trained with a multi-objective perceptual loss. Evaluation on three paired test sets with nine systems and paired significance testing leads to four findings.

First, a model evaluated with a checkpoint trained on the matching corpus is the strongest by a wide and statistically significant margin. The A2B network leads on every metric on both the Zhu and the UrbanSounds sets, including the perceptual BINAQUAL score.

Second, performance is highly sensitive to the training and test domain. On the A2B 2-MP set, for which no matched A2B or Zhu checkpoint is available, both external baselines fail, while the in-domain Proposed CRM and Wave-U-Net lead. Mixup augmentation significantly improves cross-domain transfer, but no single variant is universally best.

Third, the Proposed CRM achieves the best spectral and spatial fidelity in the cross-domain setting—the lowest LSD, ILD error, IPD error, and the highest STOI on the A2B 2-MP set—while the Wave-U-Net achieves a marginally higher SI-SDR. The perceptual localization metric further reveals that high SI-SDR does not by itself imply faithful preservation of localization cues, which points to the need for spatial evaluation measures.

Fourth, accuracy and efficiency are not aligned. The Proposed CRM offers the best balance for on-device use, combining a 3.89 M parameter footprint, real-time CPU operation (RTF $= 0.054$), and strong spectral and spatial scores, whereas the most accurate matched system (A2B) is the only one slower than real time on CPU. Future work will investigate individualized fine-tuning, higher-order Ambisonics input, and subjective listening tests.

FUNDING

This work was supported by the Polish national research budget, under the Excellence Initiative – Research University grant number 20/1/2023/IDUB/I3b/Ag.

References

[1] Zotter, F. and Frank, M., *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, volume 19 of *Springer Topics in Signal Processing*, Springer International Publishing, Cham, 2019, ISBN 978-3-030-17206-0 978-3-030-17207-7, doi: 10.1007/978-3-030-17207-7.

[2] Bernschütz, B., "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU100," in *Fortschritte der Akustik, AIA-DAGA 2013*, pp. 592–595, Merano, Italy, 2013.

[3] Zhu, Y., Kong, Q., Shi, J., Liu, S., Ye, X., Wang, J.-C., Shan, H., and Zhang, J., "End-to-End Paired Ambisonic-Binaural Audio Rendering," *IEEE/CAA Journal of Automatica Sinica*, 11(2), pp. 502–513, 2024, ISSN 2329-9266, 2329-9274, doi: 10.1109/JAS.2023.123969.

[4] Gebru, I. D., Keebler, T., Sandakly, J., Krenn, S., Marković, D., Buffalini, J., Hassel, S., and Richard, A., "A2B: Neural Rendering of Ambisonic Recordings to Binaural," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, Hyderabad, India, 2025, ISBN 979-8-3503-6874-1, doi: 10.1109/ICASSP49660.2025.10890507.

[5] Qiao, Y., Kothapally, V., Yu, M., and Yu, D., "Neural Ambisonic Encoding For Multi-Speaker Scenarios Using A Circular Microphone Array," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, Hyderabad, India, 2025, ISBN 979-8-3503-6874-1, doi: 10.1109/ICASSP49660.2025.10890048.

[6] Nawfal, I., Manias, S. D., Souden, M., Merimaa, J., Atkins, J., McMullin, E., Pirhosseinloo, S., and Phillips, D., "Ambisonics Super-Resolution Using A Waveform-Domain Neural Network," 2025, doi: 10.48550/ARXIV.2508.00240.

[7] Pulkki, V., "Spatial Sound Reproduction with Directional Audio Coding," *Journal of the Audio Engineering Society*, 55(6), pp. 503–516, 2007, ISSN 1549-4950.

[8] Yamamoto, R., Song, E., and Kim, J.-M., "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203, IEEE, Barcelona, Spain, 2020, ISBN 978-1-5090-6631-5, doi: 10.1109/ICASSP40776.2020.9053795.

[9] Rouard, S., Massa, F., and Défossez, A., "Hybrid Transformers for Music Source Separation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, Rhodes Island, Greece, 2023, ISBN 978-1-7281-6327-7, doi: 10.1109/ICASSP49357.2023.10096956.

[10] Williamson, D. S., Wang, Y., and Wang, D., "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3), pp. 483–492, 2016, ISSN 2329-9290, 2329-9304, doi: 10.1109/TASLP.2015.2512042.

[11] Roux, J. L., Wisdom, S., Erdogan, H., and Hershey, J. R., "SDR – Half-baked or Well Done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, IEEE, Brighton, UK, 2019, ISBN 978-1-4799-8131-1, doi: 10.1109/ICASSP.2019.8683855.

[12] Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S., "CBAM: Convolutional Block Attention Module," in *Computer Vision – ECCV 2018*, pp. 3–19, Springer International Publishing, Cham, 2018, ISBN 978-3-030-01234-2.

[13] Mróz, B. and Zaporowski, S., "Higher-Order Ambisonics Recording Corpus," 2026, doi: 10.34808/5xe2-ah94, dataset deposit, CC BY-NC-SA 4.0.

[14] De Coensel, B., Sun, K., and Botteldooren, D., "Urban Soundscapes of the World: selection and reproduction of urban acoustic environments with soundscape in mind," in *Proceedings of the 46th International Congress and Exposition on Noise Control Engineering (Inter-noise)*, Hong Kong, 2017.

[15] Stoller, D., Ewert, S., and Dixon, S., "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[16] Shariat Panah, D., Barry, D., Ragano, A., Skoglund, J., and Hines, A., "BINAQUAL: A Full-Reference Objective Localization Similarity Metric for Binaural Audio," *Journal of the Audio Engineering Society*, 74(1/2), pp. 20–34, 2026, ISSN 1549-4950, doi: 10.17743/jaes.2022.0243.