FiPA-SR – FiLM-Conditioned Perceptually Informed Audio Super-Resolution

Wallace Abreu and Luiz W. P. Biscainho

***Abstract*— Audio bandwidth extension aims to reconstruct missing high-frequency content from bandlimited signals. This paper proposes FiPA-SR, a GAN-based perceptual architecture capable of handling different input bandwidths within a single model. Building upon the previous AEROMamba$_{\mathrm{P}}$ framework, the proposed model incorporates FiLM layers to adapt the reconstruction process according to the respective bandwidth. Experiments on the MUSDB dataset show that FiPA-SR outperforms the state-of-the-art AudioSR model across 8, 20, and 32 kHz input sampling rates. Moreover, the proposed architecture uses approximately 3× less GPU memory and performs inference more than 60× faster than the diffusion-based baseline.**

***Keywords*— audio super-resolution, bandwidth extension, deep learning.**

I. Introduction

Audio bandwidth extension, also known as audio super-resolution, refers to the reconstruction of missing high-frequency content of a bandlimited signal [1]. The suppression of high-frequency components often produces muffled or unnatural audio, motivating the development of techniques capable of reconstructing the upper spectrum from degraded inputs.

In several scenarios involving audio transmission and storage, bandwidth limitations are commonly introduced by communication systems [2], lossy compression [3] or media degradation [4]. For technologies that prioritize intelligibility [5], such as telephony, moderate bandwidth limitations might be acceptable. However, in the case of music, high-fidelity systems must cover at least the human auditory range of 20 Hz to 20 kHz for tones [6].

Classical bandwidth extension methods relied on signal processing techniques such as nonlinear devices with linear filtering [1], source-filter modeling [7], codebook mapping [8], and spectral-band replication [9]. More recently, deep learning approaches have become dominant in audio super-resolution, with the state-of-the-art solutions using generative approaches such as generative adversarial networks (GANs) [11], [12], diffusion models [13], [14] and latent bridge models [15].

Although diffusion-based methods currently achieve state-of-the-art perceptual quality, their iterative sampling procedures [16] impose substantial computational costs during inference, making real-time or low-latency deployment challenging. When faster inference is needed, model distillation is a required step [17].

In contrast, GAN-based approaches allow direct signal generation in a single forward pass, resulting in significantly lower inference complexity [18]. Motivated by this, recent work introduced AEROMamba$_{\mathrm{P}}$ [12], a U-Net [19] based super-resolution architecture employing Mamba [20], a selective state space model designed for efficient sequence modeling, together with a differentiable Perceptual Audio Quality Measure (PAQM) [26] loss for perceptually motivated training.

However, unlike models such as AudioSR [13], which are capable of processing signals with multiple input sampling frequencies, AEROMamba$_{\mathrm{P}}$ was designed for a single bandwidth configuration per training. To address this limitation, this work proposes FiPA-SR (FiLM-Conditioned Perceptual Audio Super-Resolution), illustrated in Figure 1, an extension of AEROMamba$_{\mathrm{P}}$ that incorporates Feature-wise Linear Modulation (FiLM) [21] layers, enabling a single model to reconstruct signals from multiple bandwidth configurations and enhance them to 44.1 kHz.

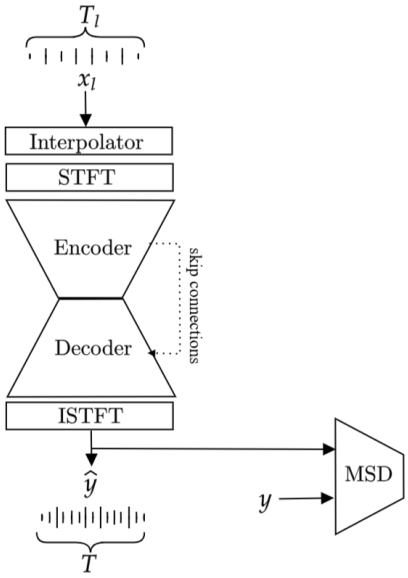

Fig. 1. General FiPA-SR architecture, including the generator and discriminator.

The motivation for this modification arises from the fact that different cutoff frequencies preserve different amounts of spectral information, directly affecting the reconstruction task. Through FiLM conditioning, the network feature maps are adaptively modulated via affine transformations controlled

Wallace Costa de Abreu, PEE/COPPE, UFRJ, Rio de Janeiro-RJ, e-mail: wallace.abreu@smt.ufrj.br; Luiz Wagner Pereira Biscainho, DEL/Poli & PEE/COPPE, UFRJ, Rio de Janeiro-RJ, e-mail: wagner@smt.ufrj.br. This work was partially supported by the Brazilian Federal Agency for Support and Evaluation of Graduate Education, CAPES (001), the National Council for Scientific and Technological Development, CNPq (306395/2025-80), and the Carlos Chagas Filho Foundation for Research Support in the State of Rio de Janeiro, FAPERJ (E-26/204.092/2022).

by a conditional vector associated with the input sampling frequency, allowing the model to adjust its behavior according to the amount of spectral content available in the degraded signal.

The proposed model is evaluated in a bandwidth extension task considering three representative bandwidth configurations somewhat associated with communication systems: 4 kHz, 10 kHz, and 16 kHz. Experimental results demonstrate that FiPA-SR consistently outperforms both AudioSR and PA-SR (an ablation version of the model without FiLM) in the objective metrics Log-Spectral Distance (LSD) [10] and Virtual Speech Quality Objective Listener (ViSQOL) [22]. Moreover, the proposed architecture utilizes in inference one-third of AudioSR GPU, and produces outputs more than 60× faster. To ensure reproducibility and allow perceptual examination, a public repository and a webpage[1] accompany this work, containing the source code, model checkpoints, and audio examples.

II. Method

The FiPA-SR architecture is illustrated in Figure 1. Similar to its predecessor, AEROMamba$_{\text{P}}$ [12], FiPA-SR is a super-resolution GAN composed of a generator based on U-Net [19] and a multi-scale discriminator. Given a signal $\boldsymbol{y} \in \mathbb{R}^T$ and its downsampled version $\boldsymbol{x}_l \in \mathbb{R}^{T_l}$, where T is the original signal length and T_l is the downsampled signal length, FiPA-SR aims to generate $\hat{\boldsymbol{y}} \approx \boldsymbol{y}$, reconstructing the lost high-frequency content of $\boldsymbol{x}_l$. In practice, the input audio is transformed into a complex spectrogram using the STFT, treating real and imaginary components as different channels, and processed by the network to estimate missing high-frequencies. The reconstructed spectrogram is then converted back to the time domain through an inverse STFT.

To handle different input sampling frequencies, a major modification introduced by FiPA-SR is the prior upsampling of the input signals to 44.1 kHz. Through this method, segments sampled at any frequency can generate spectrograms with the same size using identical parameters. The low-resolution spectrograms can then be mapped by the model into a high-frequency version. This upsampling procedure employs an interpolation biquad filter of length 128 to minimize reconstruction artifacts.

A further enhancement to the model are FiLM layers, introduced at the entrance of the residual blocks, after layer normalization (LN). The FiLM layers apply an affine transformation to an input tensor $\boldsymbol{x}$ via a pair of scaling and shifting factors $\boldsymbol{\gamma}(\boldsymbol{c})$, $\boldsymbol{\beta}(\boldsymbol{c})$ conditioned by a vector $\boldsymbol{c}$. In this setting, $\boldsymbol{c}$ contains the normalized input sampling frequencies for each sample in the batch. Formally, for a batch of size B, a vector $\boldsymbol{c} = [c_1, c_2, \ldots, c_B]^T$ is assigned, where $c_i = \frac{f_s^{(i)}}{44100}$, $i = 1, 2, \ldots, B$. Since the accepted sampling frequencies are between 0 and 44100, then $0 \leq c_i \leq 1$.

In Figure 2, the complete encoder architecture is detailed. Its core comprises two identical residual blocks composed of Snake [23] activation function, LN, Conv1D, Mamba and GLU. They are preceded by 1D convolutional layers followed by LN, the FiLM layer and a GELU. The decoder is symmetrical to the encoder and shares skip connections with the encoder layers. For adversarial training, FiPA-SR employs a MelGAN [24] multi-scale discriminator, an ensemble of three independent discriminators that operate directly on the waveform at different scales.

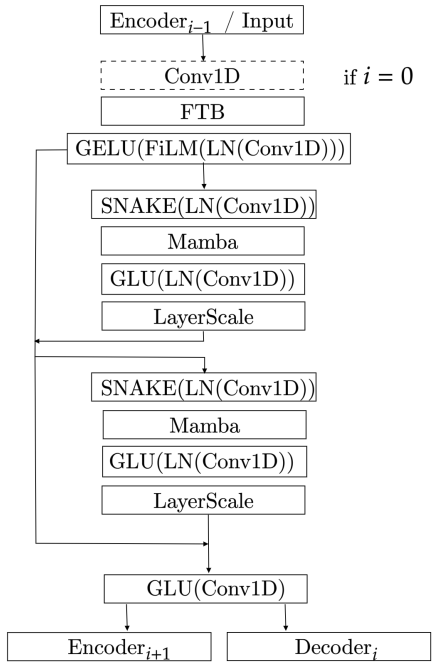

Fig. 2. FiPA-SR encoder layer architecture.

The proposed model, as in [12], is trained using a composite generator loss defined as

$$L_{\mathcal{G}} = L_{\text{adv}} + L_{\text{rec}} + \lambda L_{\text{fmap}} - \gamma L_{\text{PAQM}}, \tag{1}$$

where L_{adv} denotes the adversarial loss, L_{rec} is the spectral reconstruction loss, L_{fmap} represents the feature matching loss, and L_{PAQM} corresponds to the PAQM score computed between the reconstructed and reference signals. The parameters λ and γ control the contribution of the feature matching and perceptual loss terms, respectively.

III. Experiments

A. Dataset

The MUSDB [25] dataset is composed of 150 full-length music tracks of different genres, along with their isolated drums, bass, vocals, and other stems, totaling approximately 10 hours in duration. The training folder contains 100 songs (around 7 hours), while the test folder contains 50 songs (around 3 hours), all of which are in WAV format, stereo mode, and sampled at 44.1 kHz. The number of tracks per genre is distributed as follows: 83 Pop/Rock, 24 Rock, 15 Singer/Songwriter, 14 Heavy Metal, 13 Pop, 10 Rap, 9 Electronic, 3 Country, 3 Jazz, and 2 Reggae tracks.

Here, as in previous works [10], [12], the test set was used in validation and test phases using independent metrics. Although using the same dataset is discouraged, this procedure was used to ensure compatibility with the baseline works.

[1] https://fipa-sr.github.io/

B. Overall Setup

In the experiment, the outputs of different models were evaluated regarding their ability to enhance low-resolution signals with varied sampling frequencies. The respective input bandwidths were selected to reflect common communication channel limitations, namely: 4 kHz, typical of a telephone channel, and 16 kHz, approximating an FM broadcasting channel, along with an intermediate value of 10 kHz.

The assessed models were FiPA-SR, PA-SR, and the pretrained AudioSR model. Both FiPA-SR variants were jointly trained to handle the three considered bandwidth configurations within a single model, while AudioSR model was originally designed to support frequencies from 2 kHz up to 48 kHz. The difference between FiPA-SR and PA-SR lies in the removal of the FiLM layer in the latter, which is used as an ablation to evaluate the contribution of FiLM conditioning to the proposed architecture.

The pretrained AudioSR model served only as a supplementary baseline due to its extensive training of 7000 hours and the fact that it was exposed, in training, to the entire MUSDB dataset, including the test set, which the authors acknowledge that will benefit AudioSR in relation to the proposed models.

C. Evaluation

The models were evaluated using objective metrics: the Log-Spectral Distance (LSD) [10] and two distinct perceptual quality metrics, PAQM and the Virtual Speech Quality Objective Listener (ViSQOL) [22].

PAQM was used as a validation metric to select the optimal model during a training run, with fixed seed and hyperparameters. The choice of PAQM as a validation metric was motivated by the availability of a vectorized implementation, which is much faster than ViSQOL, and also for its perceptual motivation, taking into account masking effects and other auditory modeling aspects [26]. In addition, ViSQOL (in audio mode) was employed to assess the quality of the processed signals on a scale from 1 to 5, in relation to the test set.

The use of GPU by AudioSR and FiPA-SR was evaluated using the 'nvidia-smi' on different audio tracks. Regarding inference speed, the time required to process 10-second segments was recorded, and the average duration for each method was calculated.

Average ViSQOL scores were tested for statistical significance and, except when explicitly stated, were considered to reject the null hypothesis of no significant difference with $p-\text{value} < 0.05$.

D. Training procedure

FiPA-SR and PA-SR were trained for the bandwidth extension task considering the sampling frequencies of 8, 20, and 32 kHz. To enable a single model to handle multiple input sampling frequencies, the MUSDB dataset was resampled to each low-resolution frequency, and all tracks were jointly used during training. After frequency interpolation, the models employed a window size of $W = 512$ and a hop length of $H = 256$.

TABLE I

SUMMARY OF THE TRAINING PARAMETERS USED IN THE EXPERIMENTS.

Parameter	**Value**
Window Size (W)	512 samples
Hop Length (H)	256 samples
Segment Length	4 s
Stride Length	4 s
Number of Epochs	105
Batch Size	8
Optimizer	Adam [27]
Learning Rate	3×10^{-4}
PAQM Loss Parameter (γ)	1
Feature Matching Parameter (λ)	100

Both models were trained for 105 epochs using the default AEROMamba$_{\text{P}}$ hyperparameters, as summarized in Table I. Training was conducted on an NVIDIA RTX 3090 GPU with a batch size of 8. Finally, the model selection was based on the convergence of the PAQM score in the test set, with convergence defined as the situation where the score varies by less than 3% over five consecutive epochs.

IV. RESULTS

Results and computational performance metrics are summarized in Tables II and III, respectively. The proposed FiPA-SR achieved superior objective performance in fidelity compared to AudioSR and PA-SR, obtaining better LSD and ViSQOL scores across the evaluated bandwidth configurations. In addition, FiPA-SR demonstrated substantially lower computational complexity, requiring less than one-third of the GPU memory consumption and performing inference on 10-s audio segments more than 60 times faster than AudioSR. These advantages can be attributed both to the nearly 100× smaller number of parameters of FiPA-SR and to the inherently faster inference process of GAN-based models compared to diffusion-based sampling methods.

The objective metrics also illustrate how FiLM layers are essential to the performance of FiPA-SR, especially at lower sampling rates. At 8 and 20 kHz, PA-SR yields lower ViSQOL scores than FiPA-SR and AudioSR, indicating that the removal of FiLM significantly degrades the model's ability to generalize across multiple input bandwidth configurations. This shows that the model relies on the conditional vectors provided by FiLM to effectively create high-frequency content. In contrast, at 32 kHz, where the bandwidth extension task is less demanding due to the reduced amount of missing high-frequencies, PA-SR is able to surpass both AudioSR and baseline. A more detailed distribution of the ViSQOL results is presented in Figure 3.

The spectrogram visualizations shown in Figure 4 help illustrate the behavior of the models. The figure depicts a 5-second pop-rock segment in its original, low-resolution, and reconstructed versions for each evaluated model. AudioSR successfully extends the bandwidth for all input configurations, which is expected given its robustness as a state-of-the-art super-resolution model and its extensive training setup, including exposure to the complete MUSDB dataset. However, unlike approaches that aim for fidelity, AudioSR is primarily

Model	8 kHz		20 kHz		32 kHz	
	ViSQOL ↑	LSD ↓	ViSQOL ↑	LSD ↓	ViSQOL ↑	LSD ↓
Low-Resolution	1.64	2.17	2.44	1.82	4.10	1.64
FiPA-SR	**2.82**	**1.24**	**3.53**	**1.04**	**4.41**	**0.68**
PA-SR	2.56	1.52	3.12	1.19	4.21	0.87
AudioSR	2.72	1.69	3.33	1.30	3.85	1.06

TABLE II

OBJECTIVE EVALUATION RESULTS FOR THE BANDWIDTH EXTENSION EXPERIMENT FOR EACH INPUT SAMPLING FREQUENCY.

TABLE III

GPU USAGE (VRAM), INFERENCE TIME FOR A 10-S SEGMENT, AND TOTAL NUMBER OF PARAMETERS FOR EACH MODEL.

Method	VRAM (MB)	Time (s)	Total Parameters
FiPA-SR	3000	0.087	19,487,758
AudioSR	14396	5.663	1,285,395,637

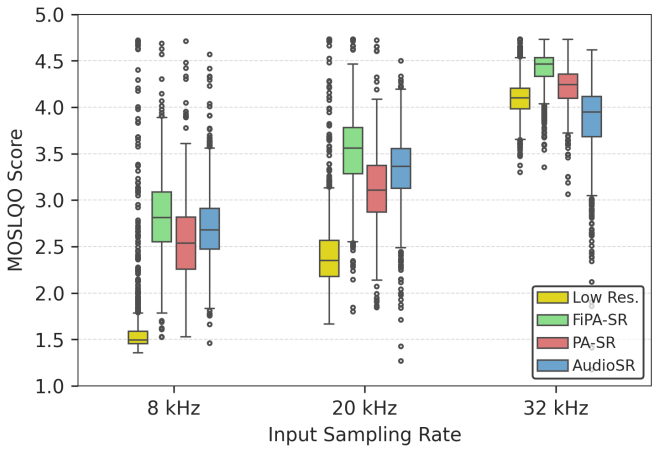

Fig. 3. Distribution of ViSQOL scores of each low-resolution and model output for varying input sampling rates.

designed to generate perceptually plausible high-frequency content rather than faithfully reconstruct the original signal. Informal listening indicates that this behavior often leads to excessive percussive transients and noticeable alterations of timbre.

The difference between FiPA-SR and PA-SR further highlights the importance of FiLM conditioning in the proposed architecture. Beyond enabling FiPA-SR to effectively handle multiple input bandwidth configurations, the FiLM layer also completely removes spectral discontinuities near the cutoff regions. In contrast, PA-SR consistently exhibits visible border artifacts in the reconstructed high-frequency regions across all evaluated bandwidths. Informal listening indicates that the perceived audio quality is consistent with the ViSQOL results, with FiPA-SR producing natural-sounding reconstructions containing few perceptible artifacts for all the three configurations.

V. CONCLUSIONS

This paper introduced FiPA-SR, a GAN-based perceptual model for audio bandwidth extension capable of handling multiple input sampling rates. Building upon the previous AEROMamba$_{\mathrm{P}}$ framework, the proposed model preserves computational efficiency while reconstructing high-frequency content across different bandwidth configurations. Experimental results demonstrated through objective metrics that FiPA-SR outperforms the state-of-the-art AudioSR model in all bandwidth settings, achieving from 0.10 to 0.56 higher MOSLQO in ViSQOL while using 3× less GPU memory and performing inference more than 60× faster.

The experiments also showed that a direct adaptation of AEROMamba to multiple bandwidth configurations, PA-SR, was not sufficient to achieve robust performance. In particular, the inclusion of FiLM layers proved essential, especially at lower sampling rates. The affine transformations controlled by the conditional vectors enabled the model to effectively generalize across different input bandwidths, while also reducing spectral discontinuities near the cutoff regions.

Future work will follow several main directions. First, the bandwidth extension experiments will be expanded to cover a broader range of sampling frequencies, from 4 kHz up to 48 kHz. In addition, subjective listening tests will be conducted to complement the objective evaluation and assess the perceptual quality of the reconstructed audio. Second, the applicability of the proposed model will be investigated in related audio processing tasks, such as encoded audio enhancement, restoration of archival recordings, and automatic equalization. Finally, from an explainability perspective, future studies will focus on understanding how different audio characteristics, such as percussive-dominant, harmonic-dominant, transient-rich, and isolated-instrument content, influence the behavior and performance of the model.

REFERENCES

[1] Larsen, E., & Aarts, R. Audio Bandwidth Extension. (John Wiley & Sons, 2004)

[2] Valin, J., & Lefebvre, R. Bandwidth Extension of Narrowband Speech for Low Bit-rate Wideband Coding. *Proceedings of the IEEE Speech Coding Workshop*. pp. 130–132 (2000)

[3] Brandenburg, K. MP3 and AAC Explained. *Proceedings of the AES 17th International Conference on High-Quality Audio Coding*. pp. 99–110 (1999)

[4] Copeland, P. Manual of Analogue Sound Restoration Techniques. (British Library, 2008)

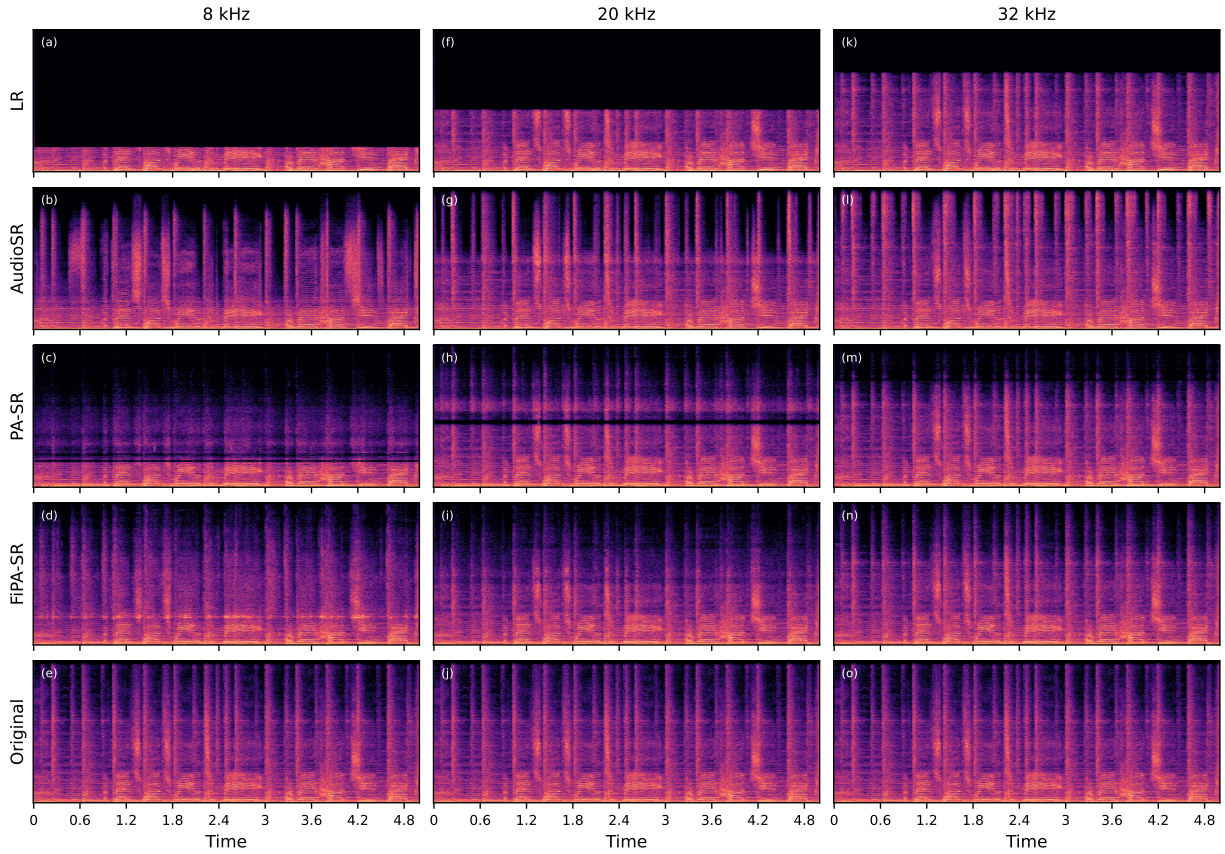

Fig. 4. Spectrograms of a 5-second audio segment for different low-resolution signals and each model output in relation to an original 44.1 kHz signal.

[5] Biscainho, L., & Nunes, L. Automatic Evaluation of Acoustically Degraded Full-band Speech. *Signals and Images: Advances and Results in Speech, Estimation, Compression, Recognition, Filtering, and Processing*. pp. 181–208 (2017)

[6] Bosi, M., & Goldberg, R. Introduction to Digital Audio Coding and Standards. (Kluwer Academic Publishers, 2002)

[7] Berisha, V., & Spanias, A. Bandwidth Extension of Audio Based on Partial Loudness Criteria. *Proceedings of the 8th IEEE Workshop on Multimedia Signal Processing*. pp. 146–149 (2006)

[8] Iser, B., & Schmidt, G. Bandwidth Extension of Telephony Speech. (Springer, 2008)

[9] Ekstrand, P. Bandwidth Extension of Audio Signals by Spectral Band Replication. *Proceedings of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio*. pp. 73–79 (2002)

[10] Mandel, M., Tal, O., & Adi, Y. AERO: Audio Super Resolution in the Spectral Domain. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. (2023)

[11] Abreu, W., & Biscainho, L. AEROMamba: An Efficient Architecture for Audio Super-Resolution Using Generative Adversarial Networks and State Space Models. *Proceedings of the 1st Latin American Music Information Retrieval Workshop*. pp. 74–79 (2024)

[12] Abreu, W., Miranda, B. V., & Biscainho, L. W. P. Efficient Audio Enhancement With a Differentiable Psychoacoustic Loss. *Journal of the Audio Engineering Society*. to appear, June 2026

[13] Liu, H., Chen, K., Tian, Q., Wang, W., & Plumbley, M. AudioSR: Versatile Audio Super-resolution at Scale. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 1076–1080 (2024)

[14] Moliner, E., Elvander, F., & Välimäki, V. Blind Audio Bandwidth Extension: A Diffusion-based Zero-shot Approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. **32**, 5092–5105 (2024)

[15] Li, C., Chen, Z., Wang, L., & Zhu, J. Audio Super-Resolution with Latent Bridge Models. *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems*. (2026)

[16] Ho, J., Jain, A., & Abbeel, P. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*. **33** (2020)

[17] Im, J., & Nam, J. FlashSR: One-step Versatile Audio Super-resolution via Diffusion Distillation. *ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 1–5 (2025)

[18] Goodfellow, I., et al. Generative Adversarial Nets. *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*. pp. 2672–2680 (2014)

[19] Ronneberger, O., Fischer, P., & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 234–241 (2015)

[20] Gu, A., & Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. Available at: https://arxiv.org/abs/2312.00752 (2024)

[21] Perez, E., Strub, F., Vries, H., Dumoulin, V., & Courville, A. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. (2018)

[22] Hines, A., Skoglund, J., Kokaram, A., & Harte, N. ViSQOL: An Objective Speech Quality Model. *EURASIP Journal on Audio, Speech, and Music Processing*. pp. 1–18 (2015)

[23] Ziyin, L., Hartwig, T., & Ueda, M. Neural Networks Fail to Learn Periodic Functions and How to Fix It. *Proceedings of the 34th International Conference on Neural Information Processing Systems*. pp. 1583–1594 (2020)

[24] Kumar, K., et al. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. pp. 14910–14921 (2019)

[25] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., & Bittner, R. MUSDB18-HQ – An Uncompressed Version of MUSDB18. Zenodo. https://doi.org/10.5281/zenodo.3338373 (2019)

[26] Beerends, J., & Stemerdink, J. A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation. *Journal of the Audio Engineering Society*. **40**, 963–978 (1992)

[27] Kingma, D., & Ba, J. Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations*. (2015)