# SAGA-SR: SEMANTICALLY AND ACOUSTICALLY GUIDED AUDIO SUPER-RESOLUTION

*Jaekwon Im* *Juhan Nam*

Graduate School of Culture Technology, KAIST, Republic of Korea

## ABSTRACT

Versatile audio super-resolution (SR) aims to predict high-frequency components from low-resolution audio across diverse domains such as speech, music, and sound effects. Existing diffusion-based SR methods often fail to produce semantically aligned outputs and struggle with consistent high-frequency reconstruction. In this paper, we propose SAGA-SR, a versatile audio SR model that combines semantic and acoustic guidance. Based on a DiT backbone trained with a flow matching objective, SAGA-SR is conditioned on text and spectral roll-off embeddings. Due to the effective guidance provided by its conditioning, SAGA-SR robustly upsamples audio from arbitrary input sampling rates between 4 kHz and 32 kHz to 44.1 kHz. Both objective and subjective evaluations show that SAGA-SR achieves state-of-the-art performance across all test cases. Sound examples and code for the proposed model are available online[1].

***Index Terms*—** audio super-resolution, bandwidth extension, flow matching, generative model

## 1. INTRODUCTION

Audio super-resolution (SR) aims to reconstruct a high-resolution audio signal from its corresponding low-resolution audio signal. To enhance listening experiences, it can be applied to diverse audio types, including historical recordings, low-bandwidth telephone audio, and audio generated by deep learning models [1]. Previous audio SR methods [2, 3] based on deep neural networks have achieved promising performance in constrained settings, including fixed upsampling ratios and restricted domains such as speech or music. However, these constraints limit their applicability in diverse and complex real-world scenarios.

To address this issue, several recent works [1, 4] have focused on versatile audio super-resolution, which is the task of upsampling general-domain audio, including music, speech, and sound effects, from varying input sampling rates to full bandwidth, such as 44.1 kHz or 48 kHz. AudioSR [1] employs a latent diffusion model (LDM) [5] with a Transformer-UNet backbone [6] to capture the complex distributions of general audio signals. FlashSR [4] employs diffusion distillation [7] to train the Student LDM using AudioSR as the Teacher LDM and proposes the SR Vocoder to further enhance AudioSR's performance.

Although previous methods have achieved notable success, there is still room for improvement. There are two key challenges hindering the performance of versatile audio SR models. First, to generate natural high-resolution audio, an audio SR model needs to capture semantic information from low-resolution input and effectively incorporate it into the reconstruction process. Previous methods often fail to predict semantically-aligned high-frequency components, resulting in unnatural artifacts, such as excessive sibilance [4]. Second, unlike speech SR, versatile audio SR handles a much broader range of audio domains, which exhibit high diversity in high-frequency energy distributions. This results in difficulties for models in consistently reconstructing high-frequency content, especially when the input has a low cutoff-frequency (e.g., 4 kHz).

In this paper, we present SAGA-SR, a versatile audio super-resolution model that leverages semantic and acoustic conditions. SAGA-SR is based on a DiT [8] backbone trained with a flow matching objective [9], incorporating two key conditions. First, inspired by recent works [10, 11] in the computer vision domain, we utilize text embeddings for semantic guidance. Specifically, we employ an audio-language model to generate text captions from audio, enabling more efficient training and inference. Second, we introduce spectral roll-off embeddings, which provide relative high-frequency energy information for both the input and target audio. Guided by both semantic and acoustic conditions, SAGA-SR can robustly upsample music, speech, and sound effects from any sampling rate between 4 kHz and 32 kHz to 44.1 kHz, and achieves state-of-the-art performance on both objective and subjective evaluations.

## 2. METHOD

Figure 1 shows the overall architecture of SAGA-SR. Let $x_h$, $x_l \in \mathbb{R}^{2\times L}$ represent the high-resolution and low-resolution audio, respectively, where $L$ is the number of samples. Each audio sample is compressed into latent representations $z_h \in \mathbb{R}^{64\times L/2048}$ and $z_l \in \mathbb{R}^{64\times L/2048}$ by the pre-trained VAE encoder from [12]. The text $c$ describes the audio content in-

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00222383).

[1] http://jakeoneijk.github.io/saga-sr-project

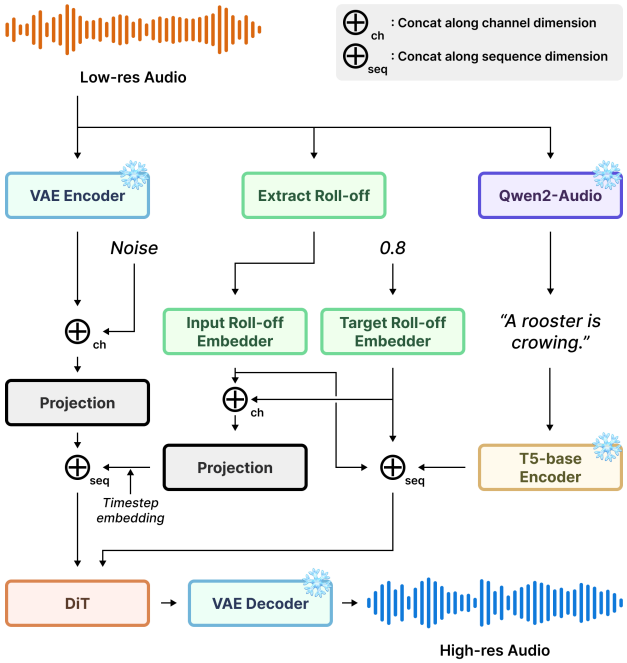


**Fig. 1**. Overview of SAGA-SR

dependent of its resolution. Roll-off frequencies $f_h$, $f_l \in \mathbb{R}$ are extracted from $x_h$ and $x_l$, respectively. DiT estimates $z_h$ from $z_l$, $c$, $f_h$, and $f_l$. The predicted latent is then converted into an audio signal by the pre-trained VAE decoder, followed by low-frequency replacement post-processing [1] to ensure consistency in low-frequency information. In Section 2.1, we present the training and inference procedures of DiT. In Section 2.2, we describe how each condition is processed to be incorporated into DiT.

### 2.1. DiT Model

Unlike previous works [1, 4] that employ a Transformer-UNet architecture, we use DiT [8], which is widely adopted in image and audio generation models. We adapt the DiT architecture proposed in [12] and train it using the conditional flow matching objective [9]. The flow matching objective regresses onto a target vector field that generates a probability path, transforming a simple distribution into an approximation of the data distribution. We use a linear interpolation path between the noise and the data as follows:

$$z_t = (1-t)\cdot z_0 + t\cdot z_1, \tag{1}$$

where $z_1 = z_h$, $z_0 \sim \mathcal{N}(0,1)$, and $t \in [0,1]$. The corresponding velocity at $z_t$ is given by

$$v_t = \frac{dz_t}{dt} = z_1 - z_0. \tag{2}$$

The training objective for DiT is defined as

$$\mathbb{E}_{t,z_0,z_1,z_l,c}\|u(z_t, z_l, c, f_h, f_l, t;\theta) - v_t\|^2, \tag{3}$$

where $\theta$ denotes the model parameters. To condition DiT on $z_l$, we concatenate $z_l$ with $z_t$ along the channel dimension. A dropout rate of 10% is applied to $z_l$, enabling classifier-free guidance [13].

At inference, we sample $z_0 \sim \mathcal{N}(0,1)$ and use an ODE solver to generate $\hat{z}_h$. In practice, we adopt the Euler sampler with a linear-quadratic t-schedule [14] and 100 inference steps. To control the influence of $z_l$ and $c$, we adopt the classifier-free guidance method introduced in [15].

$$\begin{aligned} u_{\text{CFG}}(z_t, z_l, c, f_h, f_l, t;\theta) &= u(z_t, \varnothing, \varnothing, f_h, f_l, t;\theta) \\ &+ s_a(u(z_t, z_l, \varnothing, f_h, f_l, t;\theta) - u(z_t, \varnothing, \varnothing, f_h, f_l, t;\theta)) \\ &+ s_t(u(z_t, z_l, c, f_h, f_l, t;\theta) - u(z_t, z_l, \varnothing, f_h, f_l, t;\theta)), \end{aligned} \tag{4}$$

where $s_a$ and $s_t$ are guidance scales, and $\varnothing$ denotes null conditioning. We empirically set $s_a = 1.4$ and $s_t = 1.2$.

### 2.2. Conditioning

**Text embedding.** Audio super-resolution models are typically trained with high-resolution audio-only data, since low-resolution audio can be simulated from it. Compared to audio-only data, audio-text data is expensive to curate at scale. Furthermore, relying on user-provided text during inference can limit practical applicability and potentially degrade generation quality. To address these issues, we employ Qwen2-Audio [16] to generate text captions from audio. During training, captions generated from high-resolution audio are used for efficiency, while at inference, they are derived from low-resolution audio. Text embeddings are extracted from the generated captions using a pretrained T5-base encoder [17] and provided to the DiT through cross-attention. A dropout rate of 10% is applied to the text embeddings.

**Spectral roll-off embedding.** We compute the roll-off frequency from the STFT spectrogram using an open-source method[2]. Instead of computing it frame-wise as in the original implementation, we sum over the time axis of the magnitude spectrogram to obtain a single roll-off frequency value for each audio sample, which is then normalized to $[0,1)$ using the min–max normalization. The normalized roll-off frequency value is projected into learnable Fourier embeddings. We extract the spectral roll-off embeddings from both low-resolution and high-resolution audio.

The spectral roll-off embeddings are conditioned into the DiT through two mechanisms. First, they are concatenated with the text embeddings along the sequence dimension before cross-attention. Second, the input and target roll-off embeddings are concatenated along the channel dimension, projected by linear layers, summed with the timestep sinusoidal embeddings [18], and then prepended to the input of DiT. The input roll-off embeddings provide the DiT with information about the cutoff frequency of the input audio, improving its

[2] https://librosa.org/doc/0.11.0/generated/librosa.feature.spectral_rolloff.html

ability to handle varying input sampling rates, while the target roll-off embeddings guide the amount of high-frequency energy to generate. During inference, the target normalized roll-off frequency serves as conditioning, enabling the user to control the high-frequency energy in the generated audio. Because it is represented as a single scalar in $[0, 1)$, it is straightforward to manipulate.

## 3. EXPERIMENTS

### 3.1. Training Dataset and Preprocessing

Our training dataset configuration and data simulation method are consistent with previous works [1, 4]. We train on the FreeSound [19][3], MedleyDB [20], MUSDB18-HQ [21], MoisesDB [22], and OpenSLR[4] speech dataset [23], with a total audio duration of around 3,800 hours. All audio was resampled to 44.1 kHz and randomly segmented into 5.94-second clips for training. To simulate low-high resolution audio pairs, we apply low-pass filtering to the high-resolution audio. The cutoff frequency is uniformly sampled between 2 kHz and 16 kHz. The low-pass filter type is randomly selected from Chebyshev, Butterworth, Bessel, and Elliptic, with the filter order chosen between 2 and 10.

### 3.2. Implementation Details

The DiT was trained for 26,000 steps using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use a batch size of 256 and a learning rate of $1.0 \times 10^{-5}$. An InverseLR scheduler [12] is applied with an inverse gamma of $10^6$, a power of 0.5, and a warmup factor of 0.99. To compute the roll-off frequency, we extract the STFT spectrogram using a Hann window of 2048 and a hop size of 512. The roll-off percentage is set to 0.985.

### 3.3. Evaluation

**Comparison.** We compare SAGA-SR against state-of-the-art models, AudioSR [1] and FlashSR [4]. The official implementations and checkpoints are used for all comparison models. In addition, we conducted an ablation study to evaluate the effectiveness of the text embedding and the spectral roll-off embedding. We train two SAGA-SR variants, one without the text embedding and another without the spectral roll-off embedding. Both models are trained under the same settings as SAGA-SR.

**Dataset.** For both objective and subjective evaluations, we adopt the VCTK test set (speech) [24], FMA-small (music) [25], and ESC50 fold-5 (sound effects) [26]. We selected 400 samples from each dataset.

**Evaluation Metrics.** For objective evaluation, each sound category is evaluated at cutoff frequencies of 4 kHz and 8 kHz. Log-Spectral Distance (LSD) is used as the evaluation metric, following previous studies [1, 4, 24]. Although the LSD metric is widely used in audio super-resolution tasks, it has been found that it does not always align with perceptual quality [1, 4]. To complement LSD, we also adopt the Fréchet Distance (FD) based on OpenL3 [27] for evaluating the music and sound effects categories. FD compares the statistics of embeddings from generated audio with those from ground-truth audio.

For subjective evaluation, a listening test was conducted with 25 participants. For all sound categories, the cutoff frequency was set to 4 kHz. During the test, participants were provided with the low-resolution input audio as a low anchor and asked to rate the perceptual quality of the outputs from each model on a scale from 1 to 5. We evaluated AudioSR, FlashSR, and the proposed SAGA-SR.

## 4. RESULTS

### 4.1. Objective Evaluation

Table 1 shows the results of the objective evaluation. SAGA-SR achieves state-of-the-art performance across all metrics and test cases, demonstrating the effectiveness of the proposed method. Compared to its variant without the spectral roll-off embedding, SAGA-SR consistently achieves better performance across all metrics and test cases. This indicates that the spectral roll-off embedding enhances the model's ability to handle varying cutoff frequencies and guides it to generate outputs with the desired high-frequency energy. When compared with its variant without the text embedding, SAGA-SR achieves superior performance in the speech SR task. In the music and sound effect SR tasks, SAGA-SR and its variant achieve comparable performance in terms of LSD. On the other hand, SAGA-SR consistently outperforms its variant in FD. These results demonstrate that, while the spectral roll-off embedding improves spectral alignment with ground-truth audio, the text embedding is essential for generating outputs that are more plausible and perceptually aligned with the reference audio.

### 4.2. Subjective Evaluation

Table 2 presents the results of the subjective evaluation. SAGA-SR achieves the highest scores across all test cases. Figure 2 shows a comparison of the STFT spectrograms for audio generated by different models. We found that AudioSR and FlashSR exhibit high variance in their outputs and often produce audio lacking sufficient high-frequency content. In contrast, SAGA-SR demonstrates better consistency in reconstructing high-frequency components, due to the explicit guidance provided by the spectral roll-off embedding. Moreover, we observed that AudioSR and FlashSR often generate audio that is not semantically aligned with the low-resolution input audio. Specifically, AudioSR tends to produce outputs

[3] https://labs.freesound.org/

[4] https://openslr.org/

**Table 1**. Objective evaluation results. Bold numbers indicate the best performance, while underlined numbers denote the second-best performance across all models. Note that low-frequency replacement post-processing was applied to the audio reconstructed by the VAE.

| Method | Speech | | Music | | | | Sound Effect | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 4kHz | 8kHz | 4kHz | | 8kHz | | 4kHz | | 8kHz | |
| | LSD ↓ | LSD ↓ | LSD ↓ | FD ↓ | LSD ↓ | FD ↓ | LSD ↓ | FD ↓ | LSD ↓ | FD ↓ |
| Unprocessed | 2.89 | 2.49 | 3.68 | 138.09 | 2.68 | 106.46 | 3.30 | 110.25 | 2.39 | 64.08 |
| VAE (recon) | 0.87 | 0.81 | 1.13 | 18.92 | 1.06 | 17.30 | 1.13 | 13.47 | 1.08 | 11.53 |
| AudioSR [1] | 1.46 | 1.26 | 2.09 | 32.52 | 1.88 | 25.93 | 1.85 | 39.69 | 1.68 | 28.54 |
| FlashSR [4] | 1.47 | 1.15 | 1.76 | 37.79 | 1.69 | 32.08 | 1.81 | 41.32 | 1.89 | 36.13 |
| SAGA-SR | **1.28** | **1.07** | 1.64 | **23.87** | **1.45** | **20.44** | 1.65 | **26.32** | **1.43** | **21.86** |
| w/o text | 1.32 | 1.11 | **1.63** | 30.14 | **1.45** | 25.34 | **1.60** | 29.00 | **1.43** | 23.94 |
| w/o roll-off | 1.57 | 1.43 | 2.16 | 35.99 | 1.78 | 23.5 | 2.19 | 33.07 | 1.75 | 22.4 |

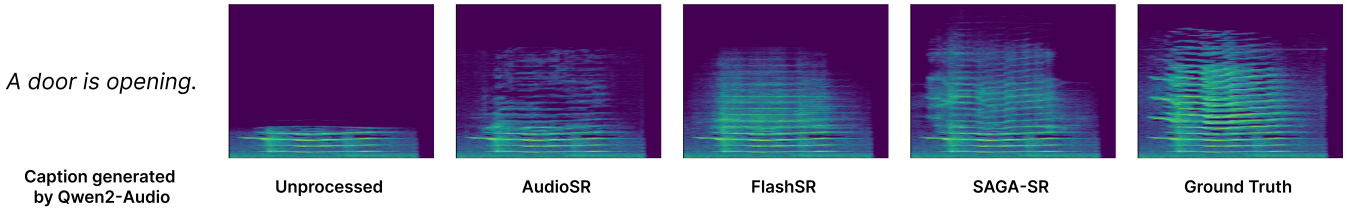


**Fig. 2**. Spectrogram of compared models.

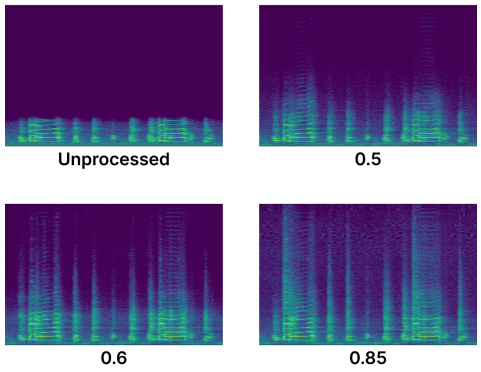


**Fig. 3**. Generated samples with varying scales of the target normalized roll-off frequency

with excessive sibilance, while FlashSR often fails to generate detailed harmonic structures, as illustrated in Figure 2. By incorporating text embeddings, SAGA-SR is able to generate semantically aligned audio with realistic harmonic detail.

As shown in Figure 3, SAGA-SR allows the user to control the high-frequency energy of the generated audio by adjusting the target normalized roll-off frequency, a single scalar condition. This control not only reduces the variance in perceptual quality but also influences acoustic properties such as timbre.

**Table 2**. Subjective evaluation results.

| Method | Speech | Music | Sound Effect |
|---|---|---|---|
| Unprocessed | 1.81 | 1.66 | 1.77 |
| Ground Truth | 4.23 | 3.93 | 4.18 |
| AudioSR [1] | 3.26 | 2.94 | 3.03 |
| FlashSR [4] | 3.45 | 3.46 | 3.34 |
| SAGA-SR | **3.70** | **3.65** | **3.88** |

## 5. CONCLUSIONS

We propose SAGA-SR, a versatile audio super-resolution model that integrates semantic and acoustic conditioning into a DiT backbone trained with a flow matching objective. Text embeddings improve semantic alignment, while spectral roll-off embeddings enhance robustness and controllability in high-frequency reconstruction. Both objective and subjective evaluations show that SAGA-SR outperforms previous methods across all tasks and metrics.

Despite its effectiveness, SAGA-SR has limitations that warrant future work. First, upsampling audio with multiple overlapping sources remains challenging. Future work could scale data and model capacity or develop improved captioning methods to accurately describe all sound sources. Second, low-frequency replacement post-processing may introduce unnatural connections between high and low frequency components. A potential direction is to develop a VAE conditioned on the low-resolution waveform to achieve smoother integration.

## 6. REFERENCES


[1] H. Liu, K. Chen, Q. Tian, W. Wang, and M. D. Plumbley, "Audiosr: Versatile audio super-resolution at scale," in *ICASSP*. IEEE, 2024, pp. 1076–1080.

[2] H. Wang and D. Wang, "Towards robust speech super-resolution," *TASLP*, vol. 29, pp. 2058–2066, 2021.

[3] S. Hu, B. Zhang, B. Liang, E. Zhao, and S. Lui, "Phase-aware music super-resolution using generative adversarial networks." in *INTERSPEECH*, 2020, pp. 4074–4078.

[4] J. Im and J. Nam, "FlashSR: One-step versatile audio super-resolution via diffusion distillation," in *ICASSP*. IEEE, 2025, pp. 1–5.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[6] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *TASLP*, vol. 32, pp. 2871–2883, 2024.

[7] C. Chadebec, O. Tasar, E. Benaroche, and B. Aubin, "Flash diffusion: Accelerating any conditional diffusion model for few steps image generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, 2025, pp. 15 686–15 695.

[8] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*, 2023, pp. 4195–4205.

[9] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *ICLR*, 2023.

[10] T. Yang, R. Wu, P. Ren, X. Xie, and L. Zhang, "Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization," in *ECCV*, 2024, pp. 74–91.

[11] X. Li, Y. Liu, S. Cao, Z. Chen, S. Zhuang, X. Chen, Y. He, Y. Wang, and Y. Qiao, "Diffvsr: Revealing an effective recipe for taming robust video super-resolution against complex degradations," in *ICCV*, 2025.

[12] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Stable audio open," in *ICASSP*. IEEE, 2025, pp. 1–5.

[13] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[14] A. Polyak, A. Zohar, A. Brown, A. Tjandra, A. Sinha, A. Lee, A. Vyas, B. Shi, C.-Y. Ma, C.-Y. Chuang *et al.*, "Movie gen: A cast of media foundation models," *arXiv*, 2024.

[15] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023, pp. 18 392–18 402.

[16] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *JMLR*, vol. 21, no. 140, pp. 1–67, 2020.

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, vol. 33, 2020, pp. 6840–6851.

[19] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *TASLP*, 2024.

[20] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive MIR research," in *ISMIR*, 2014, pp. 155–160.

[21] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The musdb18 corpus for music separation," 2017.

[22] I. Pereira, F. Araújo, F. Korzeniowski, and R. Vogl, "Moisesdb: A dataset for source separation beyond 4-stems," in *ISMIR*, 2023, pp. 619–626.

[23] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: Toward general speech restoration with neural vocoder," *arXiv*, 2021.

[24] H. Liu, W. Choi, X. Liu, Q. Kong, Q. Tian, and D. Wang, "Neural vocoder is all you need for speech superresolution," in *INTERSPEECH*, 2022, p. 4227–4231.

[25] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *ISMIR*, 2017, pp. 316–323.

[26] K. J. Piczak, "ESC: dataset for environmental sound classification," in *ACM*, 2015, pp. 1015–1018.

[27] A. L. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP*. IEEE, 2019, pp. 3852–3856.