



# MelRe: Vision-Based Mel-Spectrogram Restoration

Kaixuan Luan<sup>1</sup>, Xiaoda Yang<sup>3</sup>, Shile Cai<sup>4</sup>, Ruofan Hu<sup>3</sup>, Minghui Fang<sup>3</sup>, Wenrui Liu<sup>3</sup>, Jialong Zuo<sup>3</sup>,  
Jiaqi Duan<sup>5</sup>, Yuhang Ma<sup>6</sup>, Junyu Lu<sup>\* 2</sup>

<sup>1</sup>Beijing University of Posts and Telecommunications, China; <sup>2</sup>China Merchants Research Institute of Advance Technology, China; <sup>3</sup>Zhejiang University, China; <sup>4</sup>Harbin Institute of Technology, China; <sup>5</sup>Qingdao University, China; <sup>6</sup>Fuxi AI Lab, NetEase, China

luan123@bupt.edu.cn, lujunyu@cmhk.com

## Abstract

With advancements in visual technology, an increasing number of visual techniques have recently been applied in other fields. Among them, mel spectrograms provide a bridge between audio features and visual models. Previous work has demonstrated that applying image processing methods to mel spectrograms is feasible. However, traditional image-based models operate at a relatively coarse level, focusing primarily on controlling texture and shape. In contrast, mel spectrograms are highly sensitive to detail, containing complex time-frequency information that requires more refined modeling. To address this, we propose MelRe, a visual model specifically designed for mel spectrograms, aimed at tackling complex fine-grained audio degradation issues from a visual perspective. MelRe addresses the need for fine-grained detail through pixel-level restoration methods and employs degradation alignment and noise simulation strategies to achieve high-precision restoration across varying levels of degradation, demonstrating exceptional restoration performance. Experimental results show that MelRe achieves a new state-of-the-art (SOTA) level in complex audio restoration tasks, highlighting its potential for high-quality audio repair. Our code and demo is available at <https://melreVBMSR.github.io>.

**Index Terms:** speech recognition, human-computer interaction, computational paralinguistics

## 1. Introduction

In recent years, Audio Restoration (AR) [1, 2] has gained significant attention as an important application in digital media processing. In real-world scenarios, audio quality is often affected by multiple types of degradation, such as noise and signal loss. Traditional audio restoration models typically focus on single distortion types, making it challenging to address complex, mixed degradations, especially fine-grained noise.

Previous work has demonstrated that it is feasible to use an image processing approach for mel spectrograms, e.g. [3] convolves mel maps and ultimately extracts features well, and [4] has verified the feasibility of using a diffusion model for audio. The mel spectrogram serves as a convenient bridge between audio features and visual models and then leverage an advanced image-based diffusion model for audio restoration. Nevertheless, a mel spectrogram is a feature representation highly sensitive to detail, where its fine time-frequency structures directly impact key audio characteristics. Yet, existing image-based diffusion models are designed primarily for visual content [5], without specialized training on audio data, limiting

their effectiveness in capturing the subtle details of mel spectrograms.

To address the above challenges, we propose MelRe, a vision-based model specifically optimized for mel spectrograms, which employs a pixel-level restoration method to achieve fine-grained restoration. To accommodate varying levels of degradation, we apply degradation alignment to align images with different degradation levels to high-quality reference images. We introduce fine-grained degradation schemes into the training data and simulate various types of degradation, enabling our model to comprehensively restore complex degradations. Additionally, MelRe utilizes audio semantic guidance and classifier-free guidance techniques to further enhance restoration effectiveness. Experimental results show that MelRe achieves SOTA performance in complex audio restoration.

## 2. Method

Traditional audio restoration methods and mainstream visual diffusion models lack the ability of fine-grained restoration, while our method is committed to solving this problem. As shown in Fig. 1, the audio is first converted into a mel spectrogram matrix, which is then preprocessed and transformed into an image format. The mel image is subsequently restored using the MelRe model and finally converted into high-quality audio. The MelRe model is specifically optimized for mel spectrograms. Specifically, to achieve finer-grained perception, we employ Pixel-level Restoration, which is detailed in Sec. 2.1. To adapt to varying levels of degradation, we employ a degradation alignment module, as detailed in Sec. 2.2. To reduce the gap between training and inference, we incorporate information from low-quality (LQ) images into the inference process, as detailed in Sec. 2.3. Additionally, we integrate audio features into the generation process and employ classifier-free guidance, as detailed in Sec. 2.4. To meet the fine-grained requirements of mel spectrograms, we propose an innovative fine-grained training scheme, as detailed in Sec. 2.5.

### 2.1. Pixel-Level Restoration

ControlNet effectively supports specific conditions, such as edges and segmentation masks. However, mel spectrograms, as feature maps, require more fine-grained control. Given a feature map  $x \in \mathbb{R}^{h \times w \times c}$  from U-Net, where  $\{h, w, c\}$  are feature height, width and channel numbers, and a skipped feature map  $y \in \mathbb{R}^{h \times w \times c}$  from ControlNet. [6] proposed a unique type of convolution layer  $Z$  called zero convolution to connect them:  $\tilde{x} = x + Z(y)$ , where  $\tilde{x}$  is the output feature map. However, merely adding the feature maps from the two networks may not convey pixel-level precise information effectively, potentially causing structural inconsistencies between the input LQ

\*Corresponding author

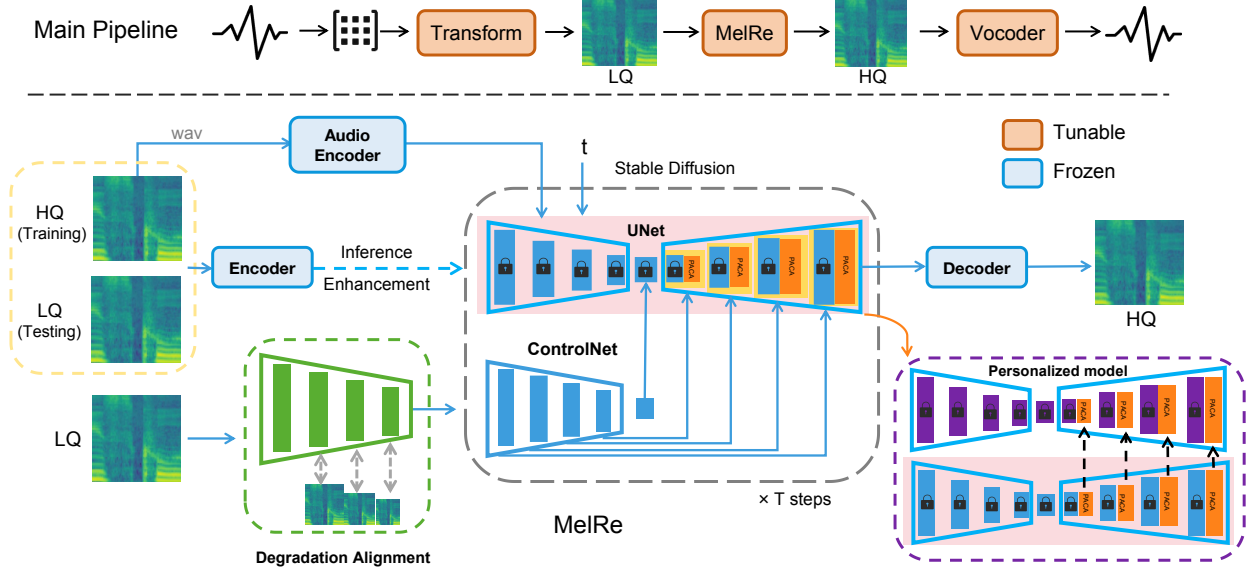


Figure 1: The pipeline of the MelRe. The upper part of the figure shows the main workflow of the model, while the lower part illustrates the structure of MelRe. In the diagram, LQ represents the low-quality image, and HQ represents the high-quality image. The encoder and decoder use the VAE architecture from Stable Diffusion. Details of the audio encoder will be discussed in the experimental section.

and output HQ images. To address this problem, we adopt a simple pixel-aware cross attention (PACA), which is inspired by [7]. We reshape  $x$  and  $y$  to  $x' \in \mathbb{R}^{h \times w \times c}$  and  $y' \in \mathbb{R}^{h \times w \times c}$ , and consider  $y'$  as the condition input. The attention matrix can be computed as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \cdot V \quad (1)$$

, where  $Q = q(x')$ ,  $K = k(y')$  and  $V = v(y')$ .

The length of the conditional feature input  $y'$  is equal to the total number of pixels in the latent feature  $x$ . Since the feature  $y'$  has not been transformed into the latent space by the encoder, it preserves the original image structure well. Therefore, our MelRe model can manage to perceive pixel-wise information from the conditional input  $y'$ .

## 2.2. Multi-Degradation Alignment

To enable the model to adapt to a wider range of image degradation levels, we introduce a multi-degradation alignment module. As shown in Fig. 1, we adopt a pyramid network to extract multi-scale feature maps with  $1/2^n$  scaled resolutions of the input LQ image, where  $n \in \{1, 2, 3\}$ . What's more, we introduce an inter-mediate supervision by employing a convolution layer to turn every single-scale feature maps into the HQ image space. We apply an L1 loss on each resolution scale to force the reconstruction at that scale to be close to the pyramid decomposition of the ground-truth image:

$$\mathcal{L}_{DA} = \sum_s \|F_{gt}^s - F_{lq}^s\|_1 \quad (2)$$

where  $F_{gt}^s$  and  $F_{lq}^s$  represent the feature embedding of the ground-truth and pyramid output at scale  $s$ .

During training, multiple alignments with HQ images are performed, creating a multi-reconstruction process that en-

courages the generation of HQ-rich features during inference. Multi-degradation simulates various levels of image degradation, enhancing the model's robustness to different degrees of noise.

## 2.3. Inference Enhancement

The noise scheduling used in stable diffusion [8] differs between training and testing, as discussed in previous works [9, 10]. During training, the noise scheduling retains some residual signal even at the final diffusion timestep  $N$ , resulting in a non-zero signal-to-noise ratio. This weakens model performance during testing, as sampling is conducted from random Gaussian noise with no signal information. To address this issue, we embed the input LQ latent into the initial random Gaussian noise at the final diffusion timestep  $N$  as compensation.

$$z_N = \sqrt{\bar{\alpha}_N} \cdot z_{lq} + \sqrt{1 - \bar{\alpha}_N} \cdot z \quad (3)$$

where  $z_N, z_{lq}, \bar{\alpha}_N, z$  are respectively the latent input at timestep  $N$ , the LQ latent, the cumulative product of  $\alpha$ , and the initial random Gaussian noise. To reduce the side effects caused by low-quality (LQ) data, we incorporate an additional Gaussian noise  $z'$  with a noise level  $\bar{\alpha}_a \in [0, 1]$  into Eq. 3 as:

$$z_N = \sqrt{\bar{\alpha}_a} \cdot (\sqrt{\bar{\alpha}_N} z_{lq} + \sqrt{1 - \bar{\alpha}_N} z) + \sqrt{1 - \bar{\alpha}_a} \cdot z' \quad (4)$$

By selecting an appropriate value for  $\bar{\alpha}_a$ , we can control the strength of the residual signal  $z_{lq}$ , allowing for a flexible balance between perception and fidelity.

## 2.4. High-Level Information

To balance mode coverage and sample quality during inference, we adopt classifier-free guidance [11]:

$$\tilde{\epsilon}(z_t, c) = \epsilon(z_t, c_{\text{neg}}) + \omega [\epsilon(z_t, c_{\text{pos}}) - \epsilon(z_t, c_{\text{neg}})] \quad (5)$$

where  $\tilde{\epsilon}(z_t, c)$  and  $\epsilon(z_t, c_{neg})$  are conditional and unconditional  $\epsilon$ -predictions [12],  $c_{pos}$  and  $c_{neg}$  are respectively the positive and negative prompts,  $z_t$  is the latent feature at step  $t$ , and  $\omega$  adjusts the guidance scale.

The unconditional  $\epsilon$ -prediction  $\epsilon(z_t, c_{neg})$  can be achieved with negative prompts. In practice, we empirically combine words like “blurry”, “low resolution”, “noisy” as negative prompts. We use an audio encoder to extract high-level semantic features of the audio,  $F_{prompt}^{pos}$ , to guide synthesis, and employ CLIP [13] to extract features of the negative prompt,  $F_{prompt}^{neg}$ . The final prompt embedding can be expressed as  $F_{prompt} = \text{Concat}(F_{prompt}^{pos}, F_{prompt}^{neg})$ .

## 2.5. MelRe Training strategy

We use RealESRGAN [14] on the training dataset to simulate various degradations, including but not limited to downsampling, blurring, and noise addition. To better simulate fine-grained noise, we apply multiple rounds of random degradation to the original image. Since the mel spectrogram is a highly detailed feature map, we further perform fine-grained noise simulation. Let the degraded image be denoted as  $I_{deg}$  and the original image as  $I_{org}$ . The fine-grained noise-simulated image  $I_{fine}$  can be obtained as follows:

$$I_{fine}^{ij} = p \cdot I_{deg}^{ij} + (1 - p) \cdot I_{org}^{ij} \quad (6)$$

where  $p \in \{0, 1\}$ , which means each pixel has a certain probability of containing noise.

Rethinking the question of how to improve the performance of the model: two main perspectives, optimizing the model structure, or improving the distribution of the dataset. Our method here starts with improving the dataset. The fine-grained noise addition process adopted increases the model’s attention to fine-grained noise, thus improving the model’s ability to handle details.

In the training process, we first obtain the latent representation  $z_0$  of a high-quality image and progressively add noise to it to produce a noisy latent  $z_t$ , where  $t$  is a randomly sampled diffusion step. Given several conditions such as diffusion step  $t$ , low-quality (LQ) input features  $F_{lq}$ , and prompt  $c$ , we train a network  $\epsilon_\theta$  to predict the noise added to the noisy latent  $z_t$ . The optimization objective is defined as:

$$\mathcal{L}_\epsilon = \mathbb{E}_{z_0, t, c, F_{lq}, \epsilon \sim \mathcal{N}(0, 1)} [|\epsilon - \epsilon_\theta(z_t, t, c, F_{lq})|_2^2]. \quad (7)$$

Then the total loss can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_\epsilon + \tau \cdot \mathcal{L}_{DA} \quad (8)$$

where  $\tau$  is a temperature coefficient used to balance the contribution of the degradation alignment loss.

Since MelRe is based on a pre-trained SD model with the pre-trained weights frozen during training, it is easy to replace the base model with a personalized model during testing, as shown Fig. 1.

## 3. Experiment

### 3.1. Dataset

VCTK [15] and LibriTTS [16] are widely used speech datasets. VCTK includes recordings from 109 native English speakers and is suitable for tasks such as speech synthesis and speech recognition. LibriTTS, derived from LibriVox audiobooks [17], contains high-quality speech samples with text annotations, making it popular for speech synthesis applications.

Unsplash2K [18], DIV2K [19, 20, 21, 22, 23], Flickr2K [24], and OST [25] are commonly used in image super-resolution and enhancement tasks. Unsplash2K and Flickr2K consist of 2000 high-quality images collected from the Unsplash and Flickr websites, offering diverse content. DIV2K includes 1000 images with 2K resolution, making it essential for image restoration tasks. OST, sourced from Open Images, is specifically designed for testing and validating super-resolution algorithms.

### 3.2. Implement Details

For training MelRe, we used the Adam optimizer with a learning rate of 5e-5, a training batch size of 4, and trained for 50 epochs, with training conditions set to gray mode. During inference, the probability of adding noise to mel-spectrogram pixels was set to 0.87, with a maximum of 2 degradation rounds. Additional noise parameters include a blur radius randomly sampled between 0.2 and 3, and a noise probability of 0.5.

Pretraining was conducted on several super-resolution datasets, including Unsplash2K, DIV2K, Flickr2K, and OST. Fine-tuning was then completed on a subset of the VCTK dataset, followed by vocoder training on the LibriTTS dataset. Our model was trained using 8 NVIDIA A800 GPUs. The vocoder was trained for 100,000 steps, with input mel-spectrograms pre-normalized.

### 3.3. Metrics

PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) [31] assess data quality by measuring differences between sets, often used for images and spectrograms. WB-PESQ [32] evaluates speech quality, focusing on intelligibility and distortion. STOI [33] measures how processing affects speech comprehension. MOS (Mean Opinion Score) is a 1–5 scale for speech quality, automated by MosNet [34]. CSIG, CBAK, and COVL are subjective scores (1–5) [35, 26], weighted by perceptual metrics. Log Spectral Distance (LSD) [36] measures spectral differences, calculating the power spectrum ratio and root mean square error. Smaller values indicate greater similarity.

### 3.4. Main Result

As shown in Tab. 1, MelRe performs exceptionally well across multiple key metrics, highlighting its technical advantages in audio restoration tasks. MelRe achieves outstanding results in PSNR and SSIM, indicating its strong capability in restoring audio details and structure. This success is due to its precise processing and feature extraction on the input mel spectrogram, which minimizes information loss during detail restoration. Additionally, MelRe’s innovative noise simulation and training strategies further enhance the quality of audio signal restoration, achieving high signal-to-noise ratios and structural similarity.

Its excellent performance on the LSD metric demonstrates MelRe’s significant advantage in suppressing spectral distortion, indicating effective error correction at the spectral level to produce audio with spectral features closer to the original. MelRe also achieves high scores in the wideband PESQ metric, showing its ability to maintain speech clarity and naturalness across various noise conditions. This performance is related to its joint denoising strategy at both spectral and waveform levels, which improves clarity while preserving natural audio quality.

In terms of STOI and MOS, MelRe also performs well. Its

Method	PSNR $\uparrow$	SSIM $\uparrow$	LSD $\downarrow$	PESQ $\uparrow$	STOI $\uparrow$	MOS $\uparrow$	COVL $\uparrow$	CBAK $\uparrow$	CSIG $\uparrow$
Resemble	19.33	0.57	9.42	2.39	0.77	3.29	2.99	2.78	3.48
AudioSR [4]	13.81	0.55	48.44	1.08	0.79	3.03	1.12	2.76	1.01
CMGAN [26]	20.75	0.66	25.12	1.07	0.75	3.10	2.85	2.58	3.53
FRCRN [27]	19.03	0.61	43.05	1.07	0.79	3.32	3.19	2.82	3.81
CleanUNet [28]	21.42	0.63	16.43	2.71	0.79	3.25	3.19	2.78	3.80
MP-SENet [29]	17.01	0.54	41.45	1.05	0.78	3.14	2.87	2.69	3.15
Voicefixer [30]	20.00	0.54	50.45	2.15	0.73	<b>3.76</b>	2.68	2.37	3.22
MelRe	<b>25.76</b>	<b>0.73</b>	<b>6.37</b>	<b>2.89</b>	<b>0.79</b>	3.20	<b>3.23</b>	<b>2.82</b>	<b>3.88</b>

Table 1: Comparison of indicators of several models.  $\uparrow$  indicates that the higher the better, and  $\downarrow$  indicates that the lower the better. MOS, COVL, CBAK, and CSIG are usually subjective indicators with a score range of (1, 5). The value ranges of SSIM and STOI are (0, 1), and the value range of PESQ is usually (-0.5, 4.5). The data in the table are processed with two decimal places reserved. Inference is conducted on a subset of the VCTK dataset, which remains unseen during training.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LSD $\downarrow$	PESQ $\uparrow$	STOI $\uparrow$	MOS $\uparrow$	COVL $\uparrow$	CBAK $\uparrow$	CSIG $\uparrow$
w/ gray&audio	25.76	0.73	6.37	2.89	0.79	3.20	3.23	2.82	3.88
w/ gray&image	25.76	0.73	6.37	2.71	0.79	3.20	3.24	2.82	3.88
w/ gray&none	25.74	0.72	6.37	2.89	0.79	3.20	3.24	2.82	3.88
w/ realisr&audio	23.97	0.68	7.30	2.60	0.79	3.13	3.12	2.60	3.79
w/ realisr&image	23.64	0.68	7.44	2.64	0.79	3.18	3.17	2.65	3.83
w/ realisr&none	24.10	0.69	7.44	2.68	0.79	3.13	3.17	2.66	3.83
w/o negprompt	25.76	0.73	6.37	2.71	0.79	3.20	3.24	2.82	3.88

Table 2: Ablation study. Gray and realisr indicate the image processing mode, specifying whether the image is handled as a grayscale or RGB image. Image denotes the use of an image feature extractor, audio refers to an audio feature extractor, none indicates no feature extraction, and ‘negprompt’ specifies whether the negative prompt is an empty string.

balanced design in feature fusion and denoising strategies enables the model to maintain high intelligibility while enhancing audio naturalness. Specific processing of the mel spectrogram likely enhances the prosody and clarity of the speech, leading to high scores in subjective quality evaluation. Furthermore, MelRe’s high score in signal integrity evaluation (CSIG) demonstrates its strong ability to preserve the structure of audio signals, indicating its robust performance in maintaining audio signal structure. It is worth noting that all experiments are conducted under the same fundamental settings with noise drawn from the same distribution

### 3.5. Ablation Study

In this section, we conduct ablation studies on the model architecture to assess the impact of each component on the model’s performance. As shown in Tab. 2, we experiment with processing modes by handling images as either grayscale or RGB full-color. For prompts, we use an image encoder to extract image features as input prompts for the model. Similarly, we perform speech recognition, feeding the recognized text into the model as a prompt through a text encoder, or alternatively using an empty string encoded as a prompt. However, experiments show that without a prompt, the generated images tended to appear overly bright. We also experiment with disabling the negative prompt. Due to the risk of model collapse when ablating certain modules, we only present the results shown in the table. Notably, treating the mel spectrogram as a color image leads to a significant performance drop, likely because the mel

spectrogram encodes information primarily in grayscale, and restoration methods designed for color images do not align well with its characteristics. Overall, the configuration shown in our structural diagram produced the best results.

## 4. Conclusion

In this work, we presented MelRe, a novel visual model specifically optimized for mel spectrogram restoration, aimed at addressing the challenges of complex pixel-level mel degradations from a visual perspective. Through pixel-level restoration module and the innovative use of degradation alignment and noise simulation, our model demonstrates exceptional capability in restoring high-quality audio even in varied and severe degradation scenarios. By integrating audio semantic guidance and classifier-free guidance techniques, our model achieves state-of-the-art performance in restoring detailed audio features that are crucial for maintaining audio quality. The experimental results show that our method not only outperforms existing methods in multiple key metrics, such as PSNR, SSIM, and LSD, but also maintains superior clarity and intelligibility in subjective evaluations, including MOS and STOI. The model’s robustness across diverse noise conditions further underscores its potential in audio restoration. Looking ahead, our approach highlights the feasibility and effectiveness of adapting advanced visual processing techniques to the audio domain, opening up new avenues for cross-modal innovations.

## 5. References

- [1] S. Godsill, P. Rayner, and O. Cappé, *Digital audio restoration*. Springer, 2002.
- [2] A. Orcalli, “On the methodologies of audio restoration,” *Journal of New Music Research*, vol. 30, no. 4, pp. 307–322, 2001.
- [3] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu *et al.*, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, vol. 12, 2016.
- [4] H. Liu, K. Chen, Q. Tian, W. Wang, and M. D. Plumbley, “Audiosr: Versatile audio super-resolution at scale,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1076–1080.
- [5] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, “Diffusion models in vision: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.
- [6] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.
- [7] P. R. X. X. Tao Yang, Rongyuan Wu and L. Zhang, “Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization,” in *The European Conference on Computer Vision (ECCV) 2024*, 2023.
- [8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
- [9] R. Girdhar, M. Singh, A. Brown, Q. Duval, S. Azadi, S. S. Rambhatla, A. Shah, X. Yin, D. Parikh, and I. Misra, “Emu video: Factorizing text-to-video generation by explicit image conditioning,” *arXiv preprint arXiv:2311.10709*, 2023.
- [10] S. Lin, B. Liu, J. Li, and X. Yang, “Common diffusion noise schedules and sample steps are flawed,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 5404–5411.
- [11] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022.
- [12] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” 2021. [Online]. Available: <https://arxiv.org/abs/2103.00020>
- [14] X. Wang, L. Xie, C. Dong, and Y. Shan, “Realesrgan: Training real-world blind super-resolution with pure synthetic data supplementary material,” *Computer Vision Foundation open access*, vol. 1, no. 2, p. 2, 2022.
- [15] J. Yamagishi, “English multi-speaker corpus for cstr voice cloning toolkit,” URL <http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html>, 2012.
- [16] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A corpus derived from librispeech for text-to-speech,” *arXiv preprint arXiv:1904.02882*, 2019.
- [17] J. Kearns, “Librivox: Free public domain audiobooks,” *Reference Reviews*, vol. 28, no. 1, pp. 7–8, 2014.
- [18] Y. Kim and D. Son, “Noise conditional flow model for learning the super-resolution,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.
- [19] E. Agustsson and R. Timofte, “Ntire 2017 challenge on single image super-resolution: Dataset and study,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- [20] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, B. Lim *et al.*, “Ntire 2017 challenge on single image super-resolution: Methods and results,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- [21] R. Timofte, S. Gu, J. Wu, L. Van Gool, L. Zhang, M.-H. Yang, M. Haris *et al.*, “Ntire 2018 challenge on single image super-resolution: Methods and results,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [22] —, “Ntire 2018 challenge on single image super-resolution: Methods and results,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [23] A. Ignatov, R. Timofte *et al.*, “Pirm challenge on perceptual image enhancement on smartphones: report,” in *European Conference on Computer Vision (ECCV) Workshops*, January 2019.
- [24] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, “Ntire 2017 challenge on single image super-resolution: Methods and results,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 114–125.
- [25] X. Wang, K. Yu, C. Dong, and C. C. Loy, “Recovering realistic texture in image super-resolution by deep spatial feature transform,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [26] S. Abdulatif, R. Cao, and B. Yang, “Cmgan: Conformer-based metric-gan for monaural speech enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [27] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, “Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9281–9285.
- [28] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, “Cleanunet 2: A hybrid speech denoising model on waveform and spectrogram,” *arXiv preprint arXiv:2309.05975*, 2023.
- [29] Y.-X. Lu, Y. Ai, and Z.-H. Ling, “Mp-senet: A speech enhancement model with parallel denoising of magnitude and phase spectra,” *arXiv preprint arXiv:2305.13686*, 2023.
- [30] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, “Voicefixer: Toward general speech restoration with neural vocoder,” *arXiv preprint arXiv:2109.13731*, 2021.
- [31] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, “scikit-image: image processing in python,” *PeerJ*, vol. 2, p. e453, 2014.
- [32] R. G. D. Miao Wang, Christoph Boeddeker and ananda seelan, “Pesq (perceptual evaluation of speech quality) wrapper for python users,” May 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6549559>
- [33] N. S. Detlefsen, J. Borovec, J. Schock, A. H. Jha, T. Koker, L. Di Liello, D. Stancl, C. Quan, M. Grechkin, and W. Falcon, “Torchmetrics-measuring reproducibility in pytorch,” *Journal of Open Source Software*, vol. 7, no. 70, p. 4101, 2022.
- [34] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, “Mosnet: Deep learning based objective assessment for voice conversion,” in *Proc. Interspeech 2019*, 2019.
- [35] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.
- [36] R. Kumar, K. Kumar, V. Anand, Y. Bengio, and A. Courville, “Nu-gan: High resolution neural upsampling with gan,” *arXiv preprint arXiv:2010.11362*, 2020.