Mutual Learning for Acoustic Matching and Dereverberation via Visual Scene-driven Diffusion

Jian Ma[1,3], Wenguan Wang[2†], Yi Yang[2], and Feng Zheng[1]

[1]Southern University of Science and Technology
[2]ReLER, CCAI, Zhejiang University [3]ReLER, University of Technology Sydney
https://hechang25.github.io/MVSD

Supplementary Materials

In the appendix, we provide the following content for a more comprehensive understanding of our method:

– § A: **Architecture Details**. We provide details of the MVSD network architecture, including layer composition, connectivity patterns, *etc*.
– § B: **Social Impacts and Limitations**.
– § C: **Qualitative Visualization** of MVSD and several competitors.

A Architecture Details

Our neural network draws inspiration from the Unet structure of Imagen [6]. Taking VAM as an example, in each step of diffusion, the controllable Unet learns to perform cross-modal generation using noisy input, clean spectrograms, and embeddings of the visual scene. As shown in Fig. A1, we divide controllable Unet into encoder and decoder with symmetric structure and both of them consist of 3 attention blocks. Skip connections [4] are employed to bridge encoder and decoder, recovering spatial information lost in downsampling. We only apply cross-modal attention [10] in the third block of the encoder and the first block of the decoder to connect visual cues and spectrograms. In self-attention block, we utilize the downsampling module [9] with a stride of 4 to rapidly reduce the size of the feature map. The feature map undergoes a size transformation in the controllable Unet ($128^2 \to 32^2 \to 8^2 \to 4^2 \to 8^2 \to 32^2 \to 128^2$). The diffusion training

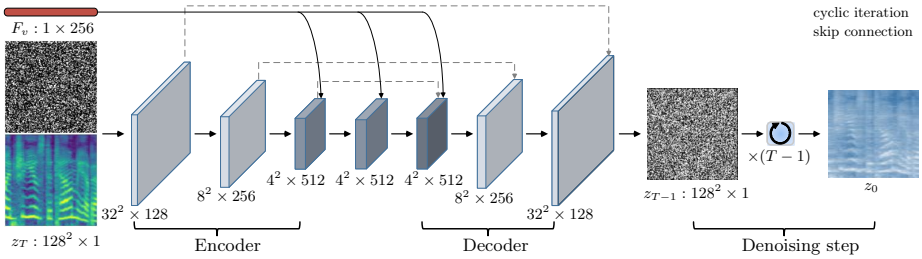

Fig. A1: Overview of the controllable Unet: MVSD utilizes a frozen visual scene encoder to encode an input RGB image into a visual feature F_v which is then mapped to a 128×128 spectrogram by Controllable Unet, facilitating auditory style transformations (§A).

process involves the following steps: starting with a sample from the data distribution, noise is gradually added over a fixed number of timesteps, creating a sequence of increasingly noisy images to reconstruct the original input. During inference, the goal is to generate samples from the learned distribution by starting with pure noise and sequentially applying the trained UNet model to denoise the image over timesteps.

B Social Impacts and Limitations

MVSD can enrich VR and AR auditory experiences with more realistic acoustics that complement the protagonist's surroundings. VAM can enhance personalized advertising, assistive technologies, and speech synthesis and recognition. Furthermore, dereverberation task can boost speech audibility across various environments, reducing reverberation for clearer communication in teleconferencing, broadcasting, and public spaces. Nevertheless, there are potential risks concerning privacy, possible misuse, and ethics, notably in diverse societal backgrounds.

While MVSD presents a promising potential, there is still scope for further exploration and investigation. Diffusion models require more time for training and sampling than GANs, posing significant challenges for real-time applications such as meetings and sound rendering, *etc*. Future efforts will focus on how to integrate methods like [8] and [5] to reduce the number of parameters and speed up diffusion models.

C Qualitative Visualization

This section showcases visualizations of qualitative results for our MVSD and competing methods. Among them, Fig. A2 and Fig. A3 depict the qualitative results of the VAM task on different datasets. Fig. A4 showcases the visualization of the generated results on SoundSpaces-Speech dataset in the dereverberation task. Fig. A5 illustrates some instances of failure cases observed on SoundSpaces-Speech dataset [1].

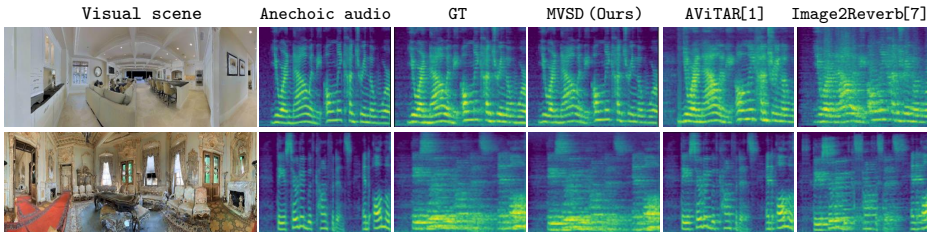

Fig. A2: Visualization results on SoundSpaces-Speech dataset in VAM task (§C).

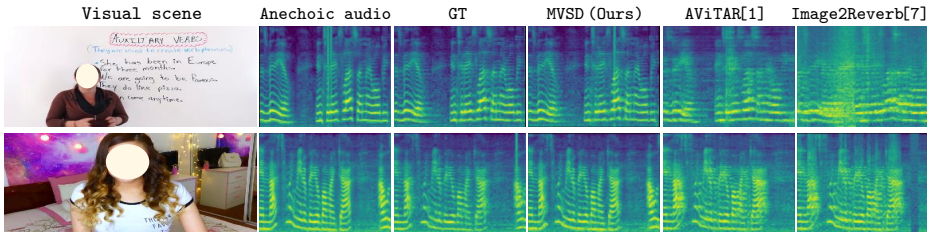

Fig. A3: Visualization results on AVSpeech dataset in VAM task (§C).

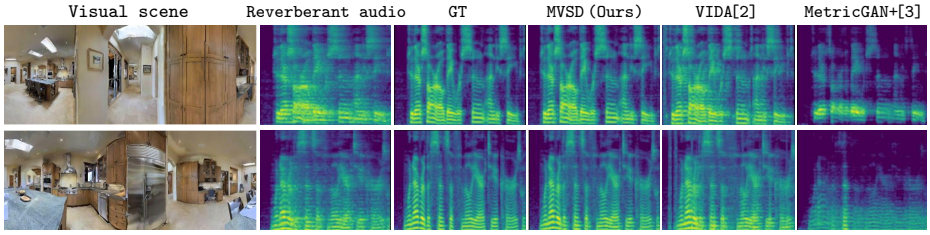

Fig. A4: Visualization on SoundSpaces-Speech dataset in dereverberation task (§C).

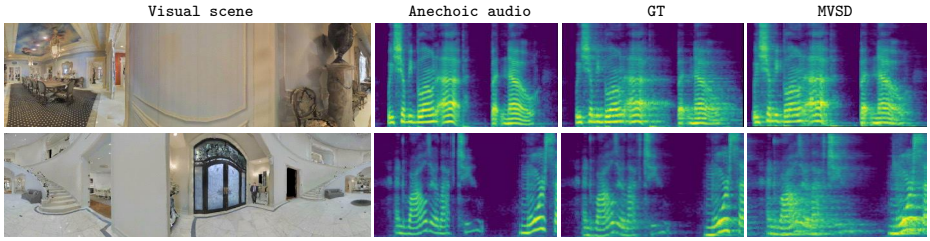

Fig. A5: Some failed samples from SoundSpaces-Speech dataset in VAM task (§C).

References

1. Chen, C., Gao, R., Calamia, P., Grauman, K.: Visual acoustic matching. In: CVPR (2022)
2. Chen, C., Sun, W., Harwath, D., Grauman, K.: Learning audio-visual dereverberation. In: ICASSP (2023)
3. Fu, S., Yu, C., Hsieh, T., Plantinga, P., Ravanelli, M., Lu, X., Tsao, Y.: Metricgan+: An improved version of metricgan for speech enhancement. In: Interspeech (2021)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
5. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. CoRR **abs/2310.04378** (2023)
6. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Lopes, R.G., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
7. Singh, N., Mentch, J., Ng, J., Beveridge, M., Drori, I.: Image2reverb: Cross-modal reverb impulse response synthesis. In: ICCV (2021)
8. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
9. Sunkara, R., Luo, T.: No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In: ECML (2022)
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)