SUPPLEMENTARY MATERIAL: AUDIO IMAGE GENERATION FOR DENOISING

Jialu Li[1] *Youshan Zhang*[2]

[1]MS in Artificial Intelligence, Yeshiva University, New York, USA
[2]Department of Computer Science and Engineering, Yeshiva University, New York, USA
yz945@cornell.edu

1. TRAINING ALGORITHM

The following algorithm describes how our WD-Mamba works to predict noise effectively.

Algorithm 1 The training process of AIGD model

Input: Noise audio signal x, clean audio signal y, and AIGD model $F(\cdot, \cdot, \theta)$.
Output: Trained image generator $F(\cdot, \cdot, \theta)$
Generate noisy audio image I_N and clean audio images I using STFT(x) and STFT(y)
while θ is not converged **do**
  for $t = T-1, \cdots, 1$ **do**
    Sample I^t using Eq. (17)
  end for
  Predict denoised image $\hat{I} = F(I^1, 1, \theta)$
  Compute gradient using the objective function in Eq. (24)
  Update θ by gradient
end while

2. RESULTS ON DNS 2020 CHALLENGE DATASET

As shown in Tab. 1, our AIGD model outperforms all other state-of-the-art models in all four metrics. Especially, the PESQ metrics are much higher than other models, which further reveals that our AIGD model achieves state-of-the-art performance in audio-denoising tasks. Therefore, the AIGD model also performs better in a real-world bird sound denoising dataset.

3. MORE RESULTS

Computation time. We also list the computation time of our AIGD model across three datasets, as shown in Tab. 2. Given our AIGD model has the T steps generation process, the computation time for each audio is increased during the training. Across three datasets, the mean computing time of training per audio is around 2.4 minutes. The major reason is due to the diffusion process and the gradient computation. Our mean inference time of the three datasets is 1.07 seconds per audio. However, compared with the best baseline methods MANNER [1], FS-CANet [2] and PtDeepLab [3] of three benchmark datasets respectively, the computation cost of our AIGD model is still in a reasonable range.

$$L = \mathcal{C}L_2 + \alpha L_{im}^{total} + (1-\alpha)L_R, \tag{1}$$

$$L_S = 1 - abs(\text{SSIM}(F(I_N), I)) = \{1 - abs(\text{SSIM}(F(I_N).real, I.real)) + 1 - abs(\text{SSIM}(F(I_N).imag, I.imag)\}/2, \tag{2}$$

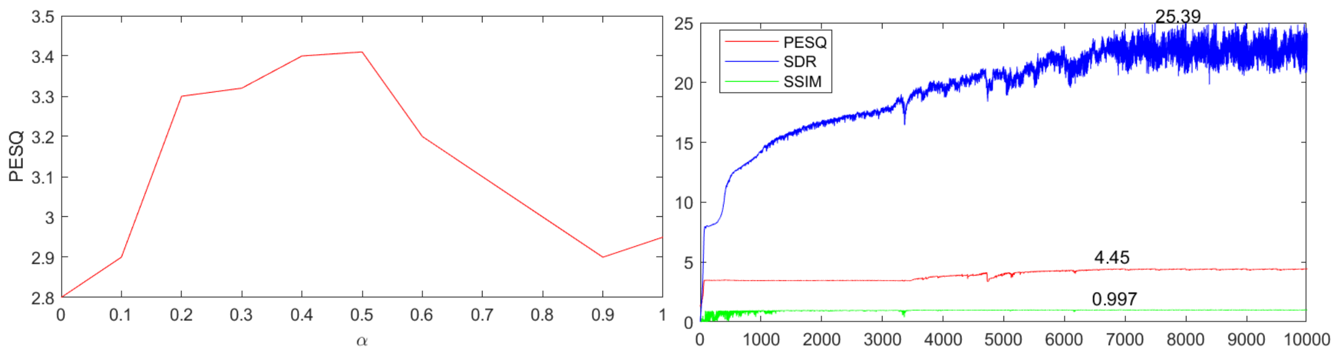
$$L_R = const_{upper} - SDR(\hat{y}, y) \tag{3}$$

Parameters analysis. In Eq. (1), α balances the image quality check loss and audio reconstruction loss. We first conduct the parameter analysis of α, which is selected from $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. “0” means that we only minimize the SDR loss, while “1” means that we only minimize the image check model. As shown in Fig. 1a, we chose $\alpha = 0.5$ as our best hyperparameter since PESQ achieves the highest value. Next, we explore the relationship among PESQ, SDR, and SSIM metrics. We defined complex absolute structural similarity loss L_S and SDR loss L_R in Eqs. (2) and (3), the lower of these two loss functions, the better the denoised results. While the higher SDR, SSIM, and PESQ, the better the results since they measure the closeness between the predictions and ground truth. As shown in Fig. 1b, with the increasing number of iterations, SSIM and PESQ converged fast and approached the highest value (1 and 4.5). Although there are some oscillations at the end of the SDR value, it is still converged, and the highest value is 25.39. Hence, we set the upper bound of SDR loss in Eq. (3) as 30.

Ablation study. To demonstrate the effectiveness of the proposed three loss functions: complex L2 norm (2), complex absolute structure similarity (S), and SDR (R), we conduct an ablation study with respect to each loss in Tab. 4. “+” means combining loss functions together. We observe that with the increasing number of loss functions, the robustness of our model keeps improving. The usefulness of loss functions is ranked as 2 > R > S. Therefore, the proposed audio image

Table 1: Comparison results on DNS 2020 challenge test dataset.

Methods	With Reverb				Without Reverb			
	WB-PESQ	NB-PESQ	STOI	SI-SDR	WB-PESQ	NB-PESQ	STOI	SI-SDR
Noisy	1.822	2.753	86.62	9.033	1.582	2.454	91.52	9.071
DTLN [4]	-	2.700	84.68	10.530	-	3.040	94.76	16.340
PoCoNet [5]	2.832	-	-	-	2.748	-	-	-
Sub-band Model [6]	2.650	3.274	90.53	14.673	2.369	3.052	94.24	16.153
FullSubNet [7]	2.969	3.473	92.62	15.750	2.777	3.305	96.11	17.290
FullSubNet+ [8]	3.218	3.666	93.84	16.810	2.982	3.504	96.69	18.340
FS-CANet [2]	3.218	3.665	93.93	16.820	3.017	3.513	96.74	18.080
AIGD	**3.381**	**3.912**	**95.02**	**18.213**	**3.351**	**4.013**	**98.31**	**20.123**



(a) Effect of different α with PESQ

(b) PESQ, SDR and SSIM changes

Fig. 1: (a): parameter analysis for α. (b): PESQ, SDR, and SSIM change with the increase of training iterations (10,000). The highest values are plotted for each line.

Table 2: Computation time (per audio) of three benchmark datasets (M: minutes, S: seconds, Voice: VoiceBank-DEMAND, DNS: DNS 2020 challenge, Bird: BirdSoundsDenoising).

Time	Voice		DNS		Bird		
	Training	Test	Training	Test	Training	Validation	Test
MANNER [1]	1.04	0.86	-	-	-	-	-
FS-CANet [2]	-	-	1.84	0.83	-	-	-
PtDeepLab [3]	-	-	-	-	0.99	0.79	0.82
AIGD	2.03	1.09	3.12	1.17	1.98	1.02	1.03

generation denoising approach is effective in improving performance, and different loss functions are helpful and important in minimizing the error between predictions and ground truth.

Reflection. From all results, we can conclude that our proposed AIGD model achieves state-of-the-art performance, which also demonstrates the superiority of the proposed architecture and novel loss functions. However, we can still observe some missing areas of generated real and imaginary images in Fig. **??**, which indicates that there is still space to further improve our model. A more effective image check loss function can be developed in future work. One weakness of our model is that it requires a high amount of GPU memory to train the model. Our AIGD model has 35M parameters, and it took around five hours per epoch, but less than one second per audio for the inference.

Table 3: Results comparisons of different methods ($F1$, IoU, and $Dice$ scores are multiplied by 100. "–" means not applicable.

Networks	Validation				Test			
	$F1$	IoU	$Dice$	SDR	$F1$	IoU	$Dice$	SDR
U^2-Net [9]	60.8	45.2	60.6	7.85	60.2	44.8	59.9	7.70
MTU-NeT [10]	69.1	56.5	69.0	8.17	68.3	55.7	68.3	7.96
Segmenter [11]	72.6	59.6	72.5	9.24	70.8	57.7	70.7	8.52
SegNet [12]	77.5	66.9	77.5	9.55	76.1	65.3	76.2	9.43
DVAD [13]	82.6	73.5	82.6	10.3	81.6	72.3	81.6	9.96
PtDeepLab [3]	83.4	75.9	83.4	10.5	83.1	75.4	83.0	10.4
AIGD	–	–	–	**11.5**	–	–	–	**10.8**

Table 4: Ablation study of different loss functions

Methods	2	S	R	2+S	S+R	2 + R	2+S+R
PESQ	3.25	2.50	2.89	3.28	2.95	3.31	3.52

4. REFERENCES

[1] Hyun Joon Park, Byung Ha Kang, Wooseok Shin, Jin Sob Kim, and Sung Won Han, "Manner: Multi-view attention network for noise erasure," in *ICASSP 2022-2022 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 7842–7846.

[2] Jun Chen, Wei Rao, Zilin Wang, Zhiyong Wu, Yannan Wang, Tao Yu, Shidong Shang, and Helen Meng, "Speech enhancement with fullband-subband cross-attention network," *arXiv preprint arXiv:2211.05432*, 2022.

[3] Junhui Li, Pu Wang, and Youshan Zhang, "Deeplabv3+ vision transformer for visual bird sound denoising," *IEEE Access*, 2023.

[4] Nils L Westhausen and Bernd T Meyer, "Dual-signal transformation lstm network for real-time noise suppression," *arXiv preprint arXiv:2005.07551*, 2020.

[5] Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy, "Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," *arXiv preprint arXiv:2008.04470*, 2020.

[6] Xiaofei Li and Radu Horaud, "Online monaural speech enhancement using delayed subband lstm," *arXiv preprint arXiv:2005.05037*, 2020.

[7] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6633–6637.

[8] Jun Chen, Zilin Wang, Deyi Tuo, Zhiyong Wu, Shiyin Kang, and Helen Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7857–7861.

[9] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern recognition*, vol. 106, pp. 107404, 2020.

[10] Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong, "Mixed transformer u-net for medical image segmentation," in *ICASSP 2022*. IEEE, 2022, pp. 2390–2394.

[11] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.

[12] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[13] Youshan Zhang and Jialu Li, "Birdsoundsdenoising: Deep visual audio denoising for bird sounds," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2248–2257.