Supplementary Material: Vision Transformer Segmentation for Visual Bird Sound Denoising

Sahil Kumar[1], *Jialu Li*[2], *Youshan Zhang*[1]

[1]Department of Artificial Intelligence and Computer Science, Yeshiva University, New York, NY, USA
[2]School of Public Policy, Cornell University, Ithaca, NY, USA.

skumar4@mail.yu.edu, jl4284@cornell.edu,youshan.zhang@yu.edu

1. More Ablation Study

Table 1: *Ablation study results comparing ViTVS variants.*

Model Variant	Validation			Test		
	IoU	$Dice$	$F1$	IoU	$Dice$	$F1$
ViTVS 1-block	37.3	59.5	59.5	36.9	58.8	59.2
ViTVS 4-block	60.3	72.6	72.7	54.5	68.4	68.4
ViTVS 5-block	68.8	76.7	76.7	67.9	76.0	76.0
ViTVS 9-block	76.8	84.7	84.7	75.2	84.0	84.0
ViTVS 11-block	78.7	87.2	87.2	78.3	86.7	88.3
ViTVS 12-block	**80.9**	**88.3**	**88.3**	**80.0**	**87.6**	**90.7**
ViTVS 15-block	78.5	86.2	86.3	78.1	85.5	85.5
ViTVS 16-block	76.1	84.5	84.5	75.3	83.8	85.9
ViTVS 17-block	76.3	84.7	84.7	75.5	83.9	85.7
ViTVS 19-block	75.4	83.9	83.9	74.7	83.4	85.4
ViTVS 20-block	80.6	88.1	88.1	79.9	87.5	88.6

Table 1 presents a comprehensive overview of the ablation study conducted on various configurations of the ViTVS model. Each row corresponds to a distinct ViTVS variant, characterized by the number of blocks in its architecture, with the table thoughtfully divided into two sections for enhanced clarity: validation and test datasets.

The metrics encompass Intersection over Union (IoU), Dice coefficient, and F1 score, which are widely employed in image segmentation tasks to gauge the accuracy and overlap of predicted segmentation masks with the ground truth. Upon scrutinizing the outcomes, a discernible progression in performance emerges with the increasing number of blocks. The ViTVS 1-block variant serves as a baseline, manifesting modest denoising capabilities. The escalation in the number of blocks, ranging from 1 to 20, consistently yields improvements in IoU, Dice, and F1 scores. Noteworthy is the ViTVS 12-block variant, standing out as the optimal configuration, attaining the highest scores across all metrics on both validation and test sets.

Of particular interest is the observation that further augmentation of the number of blocks beyond 12, as evidenced in the 15, 16, 17, 19, and 20-block variants, does not uniformly enhance performance. This implies a potential saturation point in the effectiveness of the model, emphasizing the crucial significance of a meticulously tuned architecture. Each variant underwent training for a maximum of 100 epochs, requiring approximately 25 to 35 hours. Throughout the training process, each epoch lasted between 15 to 20 minutes for models with up to 20 blocks. However, for models exceeding 20 blocks, each epoch extended to around 28 hours, prompting us to restrict the training duration to 20 epochs.

In summary, the ablation study furnishes valuable insights into the influence of block configuration on ViTVS's denoising prowess. These insights are pivotal for researchers and practitioners alike, facilitating the judicious selection of an optimal ViTVS variant for tasks pertaining to bird sound denoising.

2. Denoised Results Comparison

The visual analysis presented in Fig. 1 provides an insightful comparison of denoising outcomes, offering a comprehensive evaluation of our denoising approach employing ViTVS. In this illustration, the transformation from the original noisy audio, representing raw audio, to the denoised version by ViTVS is vividly showcased. This denoised audio is juxtaposed with the masked labeled ground truth, establishing a benchmark for evaluation.

Upon dissecting the components within Fig. 1, we discern the initial state of the audio, referred to as the original noisy audio (*Raw Audio*), encapsulating noise and undesirable elements. The ground truth is the labeled denoised reference signal, delineating the ideal denoised outcome targeted by ViTVS. Additionally, the denoised audio (*Model Output*) is the result generated by ViTVS, aiming for a close resemblance to the masked labeled ground truth.

The pivotal insight derived from the comparative analysis lies in the remarkable proximity between ViTVS's denoised audio and the labeled ground truth. In contrast to alternative denoising models, ViTVS demonstrates superior performance. Qualitatively ranking the models based on the observed results in the figure, the order from best to worst is ViTVS, Pt-DeepLab [1], DVAD [2], MAnet [3], Unet++ [4], and FPN [5]. This ranking underscores ViTVS's exceptional performance in audio denoising, particularly within the context of BirdSound-Denoising datasets.

This comparative assessment serves as a testament to the efficacy of ViTVS in achieving denoising objectives and highlights its significance in the realm of BirdSoundDenoising applications.

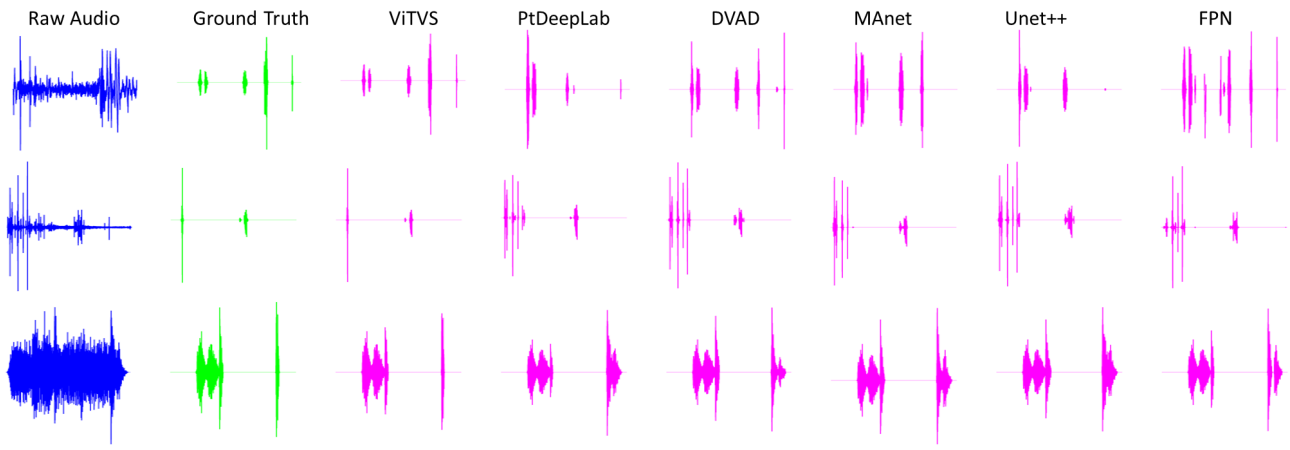

Figure 1: *Comparative analysis of denoising outcomes. The raw audio is noisy audio, and the ground truth is the clean audio. Our ViTVS shows better-denoised audios than the other five methods.*

3. References

[1] J. Li, P. Wang, and Y. Zhang, "Deeplabv3+ vision transformer for visual bird sound denoising," *IEEE Access*, 2023.

[2] Y. Zhang and J. 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.

[3] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 1–13, 2022. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2021.3093977

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," 2018.

[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2017.