

1 Appendix

1.1 Models architecture details

Table 1 shows the architecture of the Unet3D model used for all denoisers in the experiments. The model consists of 3 downsample blocks, 2 middle blocks, and 3 upsample blocks. Each block consists of a convolutional layer, followed by a spatial linear attention layer, and an attention layer for same pixel position. The number of parameters in each layer is also shown in the table. The input shape is $(B, 3, K, H, W)$, where B is the batch size, K is the number of frames, H is the height, and W is the width. The output shape is $(B, 3, K, H, W)$, where 3 is the number of channels (RGB). The number of channels in the input and output is 3, as the input and output are RGB images. The number of channels in the intermediate layers is 32, 64, 128, 256, and 512, respectively. The number of channels in the output of the upsample blocks is 128, 64, and 32, respectively. The final convolutional layer outputs the denoised image with the same shape as the input. The number of parameters in the model is 1,036,195. The model is implemented using the PyTorch library.

1.2 Supplementary Videos

In the supplementary material, we provide several videos to demonstrate the effectiveness of the proposed method. including the following:

- **Compare with other methods:** We provide four videos to compare the proposed method with other methods for each dataset. The videos show the segmentation map, classifier-free condition, our method, and the ground truth.
- **Effect of the denoising step:** We provide four videos to show the effect of the start denoising step for each dataset. The videos show the segmenation and result for each starting denoising step.

Table 1: Description of Layers in Unet3D Model, used for all denoisers in the experiments.

Block	Input Shape	Output Shape	Layer	Parameter Count
Initial Convolution	(B, 3, K, H, W)	(B, 32, K, H, W)	Conv3D	672
			Conv3D	2112
Downsample 1	(B, 32, K, H, W)	(B, 64, K, H/2, W/2)	Conv3D	2112
			SpatialLinearAttention	256
			Attention	576
			Conv3D	16512
Downsample 2	(B, 64, K/2, H/2, W/2)	(B, 128, K/4, H/4, W/4)	Conv3D	8320
			SpatialLinearAttention	512
			Attention	1152
			Conv3D	66048
Downsample 3	(B, 128, K/4, H/4, W/4)	(B, 256, K/8, H/8, W/8)	Conv3D	33024
			SpatialLinearAttention	1024
			Attention	2304
			Conv3D	264192
Middle Block 1	(B, 256, K/8, H/8, W/8)	(B, 256, K/8, H/8, W/8)	Conv3D	33024
			SpatialLinearAttention	1024
			Attention	2304
			Conv3D	264192
Middle Block 2	(B, 256, K/8, H/8, W/8)	(B, 256, K/8, H/8, W/8)	Conv3D	33024
			SpatialLinearAttention	1024
			Attention	2304
			Conv3D	264192
Upsample 1	(B, 512, K/4, H/4, W/4)	(B, 128, K/4, H/4, W/4)	Conv3D	8400
			SpatialLinearAttention	512
			Attention	1152
			Conv3DTranspose	264192
Upsample 2	(B, 256, K/2, H/2, W/2)	(B, 64, K/2, H/2, W/2)	Conv3D	2112
			SpatialLinearAttention	256
			Attention	576
			Conv3DTranspose	66048
Upsample 3	(B, 128, K, H, W)	(B, 32, K, H, W)	Conv3D	672
			SpatialLinearAttention	64
			Attention	144
			Conv3DTranspose	16512
Final Conv	(B, 64, K, H, W)	(B, 3, K, H, W)	Conv3D	195
			Conv3D	672