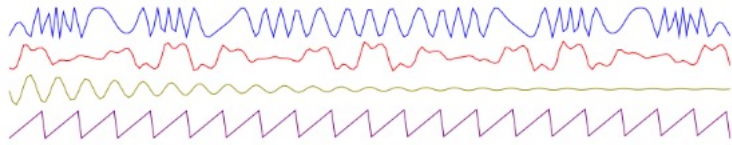
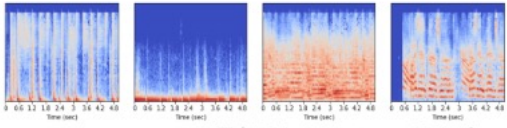


ISTFT



T/1024 time steps, 2048 freq.

T time steps.

$\text{ZDecoder}_1(C_{in} = 48, C_{out} = 4 \cdot 2 \cdot 2)$

$\text{TDecoder}_1(C_{in} = 48, C_{out} = 4 \cdot 2)$

T/1024 time steps, 512 freq.

T/4 time steps

$\text{ZDecoder}_2(C_{in} = 96, C_{out} = 48)$

$\text{TDecoder}_2(C_{in} = 96, C_{out} = 48)$

T/1024 time steps, 128 freq.

T/16 time steps

$\text{ZDecoder}_3(C_{in} = 192, C_{out} = 96)$

$\text{TDecoder}_3(C_{in} = 192, C_{out} = 96)$

T/1024 time steps, 32 freq.

T/64 time steps

$\text{ZDecoder}_4(C_{in} = 384, C_{out} = 192)$

$\text{TDecoder}_4(C_{in} = 384, C_{out} = 192)$

T/1024 time steps, 8 freq.

T/256 time steps

Cross-Domain Transformer Encoder

T/1024 time steps, 8 freq.

T/256 time steps

$\text{ZEncoder}_4(C_{in} = 192, C_{out} = 384)$

$\text{TEncoder}_4(C_{in} = 192, C_{out} = 384)$

T/1024 time steps, 32 freq.

T/64 time steps

$\text{ZEncoder}_3(C_{in} = 96, C_{out} = 192)$

$\text{TEncoder}_3(C_{in} = 96, C_{out} = 192)$

T/1024 time steps, 128 freq.

T/16 time steps

$\text{ZEncoder}_2(C_{in} = 48, C_{out} = 96)$

$\text{TEncoder}_2(C_{in} = 48, C_{out} = 96)$

T/1024 time steps, 512 freq.

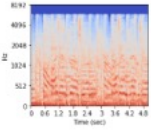
T/4 time steps

$\text{ZEncoder}_1(C_{in} = 2 \cdot 2, C_{out} = 48)$

$\text{TEncoder}_1(C_{in} = 2, C_{out} = 48)$

T/1024 time steps 2048 freq.

T time steps



STFT