RSTT: Real-time Spatial Temporal Transformer for Space-Time Video Super-Resolution Supplementary Materials

A. Attention visualization

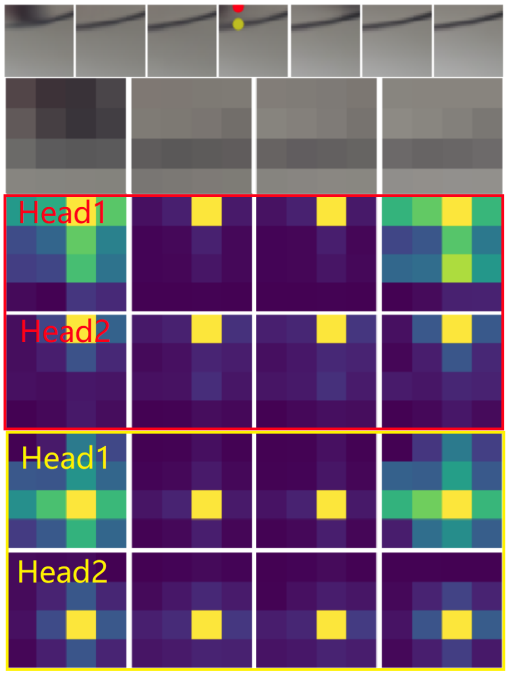

Figure 1. Windowed 2-Head attention of the last decoder.

To better understand the mechanism of RSTT, We show windowed 2-Head cross attentions of the last decoder on a sequence in Vimeo-Fast at two different locations (see Figure 1). Images bounded within red (yellow) box correspond to the red (yellow) dot highlighted in the outputs. We observe that the learnt attentions generally capture the local image structures in the finest level of detail.

B. Effectiveness of Multi-head Cross Attention

RSTT employs Multi-head Cross Attention (MCA) to generate features in Decoders based on corresponding Encoders and Queries. To further investigate the effectiveness of this design, we replace MCA in RSTT-S with other ways of information combinations: concatenation and addition. The performance comparisons are shown in Table 1 and 2.

Decoder	Vid4		Vimeo-Fast	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
MCA	26.29	0.7941	36.58	0.9381
Concat	26.18	0.7879	36.29	0.9346
Add	26.13	0.7865	36.25	0.9340

Table 1. **Quantitative comparisons on Vid4 and Vimeo-Fast datasets between MCA and other information fusion methods.** PSNR and SSIM are computed on Y channel only.

Decoder	Vimeo-Medium		Vimeo-Slow	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
MCA	35.43	0.9358	33.30	0.9123
Concat	35.20	0.9332	33.15	0.9099
Add	35.15	0.9327	33.07	0.9090

Table 2. **Quantitative comparisons on Vimeo-Medium and Vimeo-Slow datasets between MCA and other information fusion methods.** PSNR and SSIM are computed on Y channel only.

Here, both Windowed-MCA and Shifted Windowed-MCA are replaced by feature concatenation or feature addition by simply using 1d convolutions to match the feature sizes of stacked queries and the encoder outputs before this information fusion. Metrics in both Table 1 and 2 clearly testify the effectiveness of MCA over feature concatenation and feature addition.