Supplementary Material: Cross-Resolution Flow Propagation for Foveated Video Super-Resolution

Eugene Lee Lien-Feng Hsu Evan Chen Chen-Yi Lee
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
{eugene.ee06g,lienfeng.ee09g,evanchen.ee06}@nctu.edu.tw, cylee@si2lab.org

1. Network Configuration

All convolutional layers before the final feature aggregator have output channel size of 32, this includes the convolutional layers embedded within the encoder $\mathcal{E}^{\mathrm{LR}}$ and within the feature aggregator blocks. All convolutional layers are paired with a LeakyReLU activation function to model non-linearity. The final feature aggregator and the fovea encoder $\mathcal{E}^{\mathrm{Fv}}$ have output channel of size 4. $\mathcal{C}_{\mathrm{fb}}$ has input channel of 8 (concatenation of foveated region features and features from feature aggregator) and output channel of 4. $\mathcal{C}_{\mathrm{out}}$ has an output channel of 3.

Flow Field Estimator. The flow field estimator $\mathcal{F}$ has an encoder-decoder structure that maps images of the current and previous time step, i.e. $\mathbf{I}_t^{\mathrm{LR}}$ and $\mathbf{I}_{t-1}^{\mathrm{LR}}$, to the flow field $\mathbf{F}_t$. To meet real-time inference latency, we construct our own flow field estimator. The flow field estimator is composed of 3 encoder blocks, 3 decoder blocks and a flow estimation block. Both encoder and decoder blocks are composed of two convolutional layers followed by ReLU activation. Average pooling of kernel size 2 is placed right after each encoding block. The flow estimation block has two convolutional layer with a ReLU activation layer in between and a tanh activation layer at its output.

2. Experiments on DCN State Vector

DCN state vectors (DSV) are introduced to retain state information that are useful in super-resolving future frames. The introduction of DSV helps in reducing the required parameters and computational cost as less features are propagated towards the upcoming feature aggregators and are stored internally as state vectors within the feature aggregator blocks. Here, we perform a study on the trade-off between the allocation of features for forward propagation or are propagated internally within each feature aggregators as DSV. We summarize the ablation study on DSV in Table 1. We can observe that the introduction of a small amount of DSV into the feature aggregator contributes to the final performance.

3. Simulating FVSR with Eye Tracker Noise

We show similar a simulation as the one shown in the main paper in Figure 1.

4. Analysis of CRFP-Fast

For CRFP to achieve real-time latency for head-mounted device, only a fixed region (720×720) is passed through the DCN blocks within the feature aggregator for fine-grained warping while the rest are forward propagated through the residual block within the feature aggregator. Using this approach, we are able to reduce the latency by a factor of 3 (latency of 14 ms per frame using RTX 3090), enabling real-time inference using our architecture. Although CRFP-Fast has low VMAF score in the main paper, it is shown to be visually pleasing in Figure 2. As pixel region far beyond the foveal acuity are not efficiently picked up by our visual system, loss in visual quality in that region doesn't not affect our visual perception of the video.

5. Video in Supplementary Materials

We show videos with the format of Figure 2 in our supplementary materials. The name formatting of the videos follows the rule σ^{T}_videoID.mp4. σ^{T} correspond to the standard deviation of the distribution of the additive Gaussian noise introduced to the foveated region.

Table 1. Performance comparison of $8\times$ FVSR evaluated using REDS4 at proposed regions using PSNR, SSIM and VMAF. Comparison using various input configuration of the feature aggregation is shown. Setting the total input channels as 32, we study the trade-off in ratio between the features from the previous feature aggregator ($\hat{\mathbf{h}}_{t-1} \oplus \mathbf{h}_t^l$) and the DSV embedded within the current feature aggregator.

Channels		Foveated Region		Past Foveated Region(s)		Whole Image		
$\hat{\mathbf{h}}_{t-1} \oplus \mathbf{h}_t^l$	DSV	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	VMAF
8	24	**42.27**	0.9835	30.29	0.8361	25.87	0.7246	67.12
16	16	42.12	0.9834	30.47	0.8424	25.96	0.7292	69.88
24	8	42.14	**0.9836**	**30.59**	**0.8455**	**26.07**	**0.7338**	**70.30**
32	0	41.31	0.9831	29.96	0.8242	25.78	0.7182	66.58

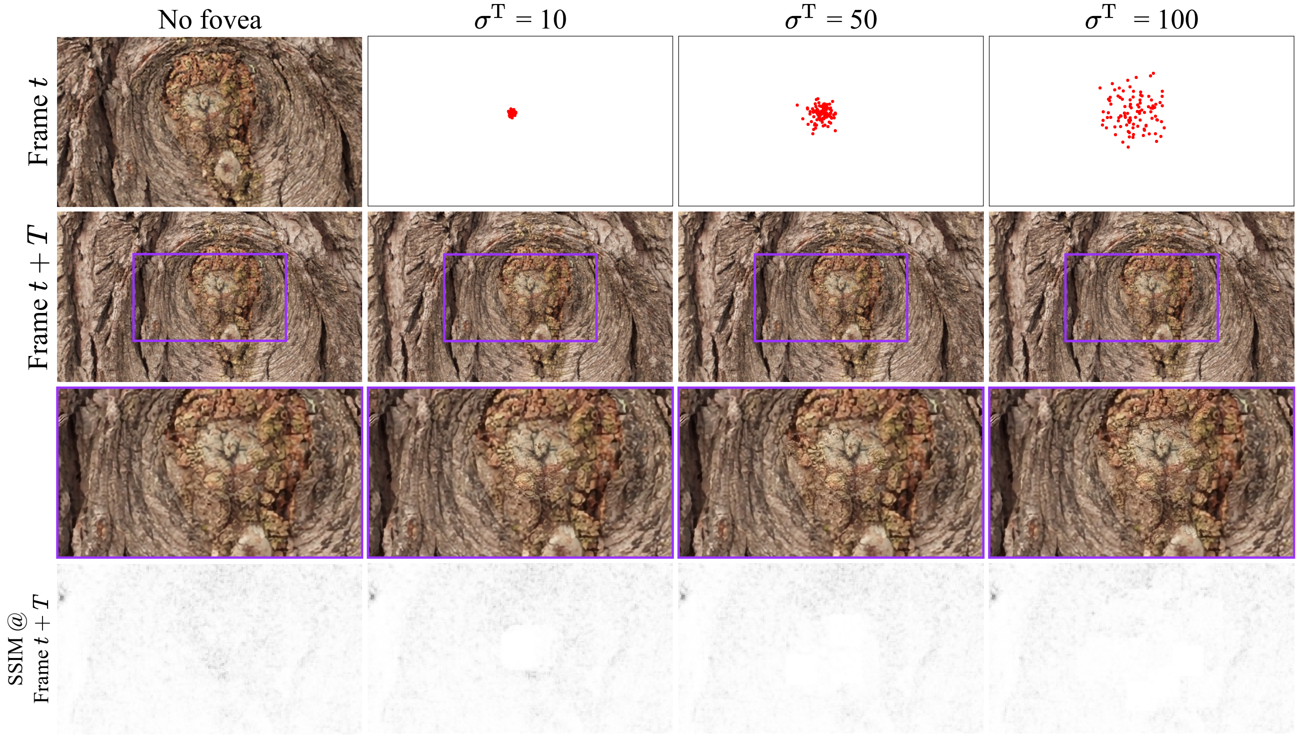
No fovea
$\sigma^{\mathrm{T}} = 10$
$\sigma^{\mathrm{T}} = 50$
$\sigma^{\mathrm{T}} = 100$
Frame t
Frame $t + T$
SSIM @ Frame $t + T$

Figure 1. Simulating the actual use case of FVSR where there is additive Gaussian noise present in an eye tracker. Various standard deviations σ^{T} are tested and results show that larger σ^{T} demonstrates the capability of the model on retaining HR context from past foveated regions. Spot the difference in details of the stripes on the log. SSIM plots are also provided to assist the reader in spotting the differences across different σ^{T}. Larger σ^{T} results in larger coverage of HR region but loses marginal detail at the center point.

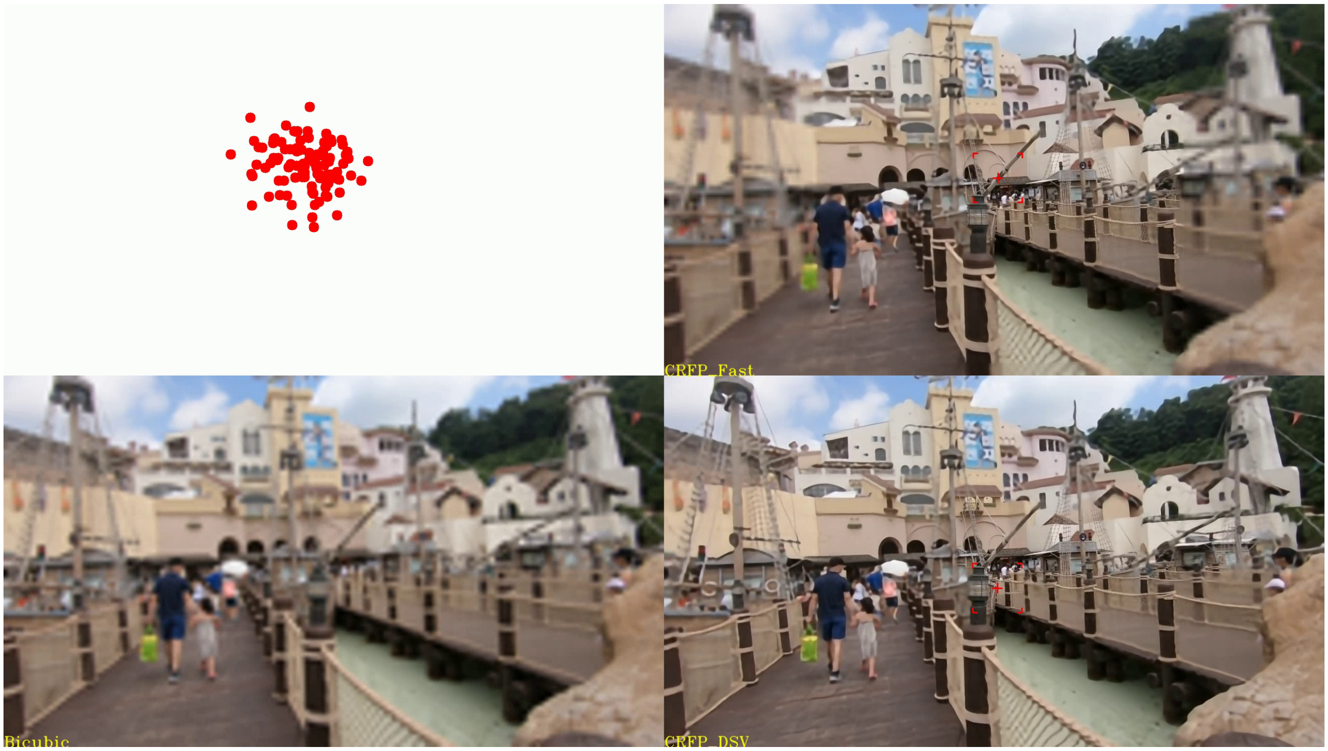

Figure 2. **Top Left**: 100 center points of foveated region; **Top Right**: CRFP-Fast after 100 frames; **Bottom Right**: CRFP + DSV after 100 frames; **Bottom Left**: Bicubic result after 100 frames. Notice that while CRFP has noticeably lower quality beyond the region (720 × 720) passed into the DCN, it is not visually perceptible if we focus on the foveated region.