

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

Abstract

The demand of high-resolution video contents has grown over the years. However, the delivery of high-resolution video is constrained by either computational resources required for rendering or network bandwidth for remote transmission. To remedy this limitation, we leverage the eye trackers found alongside existing augmented and virtual reality headsets. We propose the application of video super-resolution (VSR) technique to fuse low-resolution context with regional high-resolution context for resource-constrained consumption of high-resolution content without perceivable drop in quality. Eye trackers provide us the gaze direction of a user, aiding us in the extraction of the regional high-resolution context. As only pixels that falls within the gaze region can be resolved by the human eye, a large amount of the delivered content is redundant as we can't perceive the difference in quality of the region beyond the observed region. To generate a visually pleasing frame from the fusion of high-resolution region and low-resolution region, we study the capability of a deep neural network of transferring the context of the observed region to other regions (low-resolution) of the current and future frames. We label this task a Foveated Video Super-Resolution (FVSR), as we need to super-resolve the low-resolution regions of current and future frames through the fusion of pixels from the gaze region. We propose Cross-Resolution Flow Propagation (CRFP) for FVSR. We train and evaluate CRFP on REDS dataset on the task of $8\times$ FVSR, i.e. a combination of $8\times$ VSR and the fusion of foveated region. Departing from the conventional evaluation of per frame quality using SSIM or PSNR, we propose the evaluation of past foveated region, measuring the capability of a model to leverage the noise present in eye trackers during FVSR. Code is made available at <https://github.com/eugenelet/CRFP>.

1. Introduction

The impact of video super-resolution (VSR) in our daily life has become more prominent in the recent years as high

quality contents can be delivered while its lower quality counterpart is rendered or stored, saving either computational or storage resources. The application of deep neural networks to the task of rendering high-resolution frames using its low-resolution sampled counterpart has brought forward substantial improvements that enables technologies like Deep Learning Super-Sampling (DLSS) [11] and Deep-Fovea [24, 51]. They deliver high quality content on a computationally-constrained platform. While existing VSR techniques are implemented on a pixel level and are able to reconstruct video content to a point that is visually pleasing, certain context that are of high-quality that is meant to be delivered might not be fully reconstructed, restricting high frequency context from being delivered, e.g. texts and fine textures. Results of VSR techniques are visually acceptable up to $4\times$ VSR, while frames generated using $8\times$ VSR have distinctive flaws that affects the overall viewing experience. We argue that VSR methods while being useful for delivering general contexts, it should open up the possibility of fusing super-resolved frames with regional high-resolution (HR) context(s), e.g. HR patches, that are crucial for the understanding of the gist of the delivered content.

With the increase in adoption of augmented and virtual reality (AR/VR) devices [12, 49, 17], the demand for high-resolution content will show similar spike. As more pixels are required for the immersive experience for AR/VR, developers are searching for effective ways to reduce the computational cost of rendering frames for AR/VR. A feasible approach is to include an eye tracker in the AR/VR headset to estimate the gaze direction of the user [7]. Frames are rendered based on the gaze direction of the user [40, 24], resulting in huge reduction in computational cost. Our work leverages the eye tracker of such devices for the task of Foveated Video Super-Resolution (FVSR). FVSR is useful if we are to transfer HR content to be viewed in real-time, especially to AR/VR devices. Transferring the HR frames at its full resolution might not be feasible at a bandwidth-constrained environment. For FVSR, only the pixels that fall in gaze region are transmitted in HR while the rest are transmitted in low-resolution (LR). This results in huge sav-

ings in bandwidth as it is empirically shown that the human eye is only able to perceive and resolve around $\sim 1\%$ of pixels in a frame [46, 15]. The main challenge of FVSR is the transfer of context from the HR to the LR region, to prevent abrupt transition in visual quality.

As prior works of VSR don't consider the fusion of LR and HR context, cross-temporal operation or alignment is performed on the feature maps of lowest spatial resolution, i.e. during the propagation and aggregation stages [4, 5]. The temporally-aggregated low spatial resolution feature maps are then upsampled using a sequence of upsampling filters to reconstruct the HR frame. To incorporate regional HR context(s) into the super-resolution pipeline, we need to have precise spatial locality for the placement of the HR region. To do so, we propose a Cross-Resolution Flow Propagation (CRFP) framework that follow existing VSR framework which sequentially performs propagation, alignment, aggregation and upsampling. To aggregate the foveated context into the super-resolution pipeline, few modifications are made. Foveated region are fed to the *Feature Aggregator* (FA) using a feedback mechanism. Multiple FAs are placed at the features of lowest resolution and a single FA is placed before an output block with features having the targeted resolution. As the spatial resolution of the feature maps of the final stage (after upsampling stages) matches the spatial resolution of the HR region, i.e. matching coordinates, the spatial fusion of both features can be precise. Leveraging VSR techniques for the construction of CRFP, we show promising results for FVSR that are not achievable by existing VSR techniques.

The closest work to our proposed research direction of FVSR is DeepFovea [24]. DeepFovea performs video inpainting given a sequence of sparse frames. As FVSR is a novel task, the only comparison we make is with the architecture we bootstrap CRFP on, BasicVSR++ [5] (SoTA in VSR), modified for the task of FVSR. Our contributions are summarized as follows:

1. We propose a new task of Foveated Video Super-Resolution (FVSR). FVSR requires an eye tracker to work and is applicable to the growing adoption of AR/VR devices for the streaming of HR video.
2. We propose a Cross-Resolution Flow Propagation (CRFP) technique for FVSR, demonstrating convincing results for FVSR.
3. To quantitatively measure the performance of FVSR, we propose the evaluation of Past Foveated Region using PSNR and SSIM to better evaluate the capability of a model to retain contexts from previous frames.

2. Background and Related Work

Visual Perception of Foveated Video. As video contents are designed to be consumed by the human eye, we can exploit how visual signal is encoded for processing at the visual system to compress our data source without inducing perceptible loss in visual quality. Curcio *et al.* [8] shows that there's a rapid decrease in the number of photoreceptors in the eye from the fovea to the periphery, also known as eccentricity. Despite the loss in spatial resolution, Rovamo *et al.* [37] shows that temporal sensitivity remains static spatially, requiring the displayed video to have smooth transition across frames. The perception of spatial detail at a certain spatial frequency (visual acuity) is limited by the density of the midget ganglion cells that provide the pathway out of the eye [26, 36]. Dacey and Paterson [9] show that there's an order of $30\times$ reduction in cell density from the fovea to periphery ($0^\circ - 40^\circ$), giving us a hint on the size of the foveated region to be cropped from the HR image. The central 5.2° region of the retina has high sensitivity, covering only 0.8% of total pixels on a regular display [46, 15]. This finding points us to the choice of cropping $\sim 1\%$ of the total pixels as the foveated region in our experiments.

Studies in [44, 48] shows that in peripheral regions, mismatch between optical, retinal and final neural sampling resolutions leads to aliasing in our peripheral vision. *Aliasing zone* is the gap between the detection and resolution thresholds [35]. Context between the detection and resolution threshold are details can be detected but not resolved whereas context within the resolution threshold can be clearly resolved and detected. The role of FVSR is to attempt to reconstruct the context of the targeted frame such that the quality of the reconstructed pixels within the aliasing zone is visually pleasing. This is measured by the outskirts of foveated region in the experimental section. Naive downsampling of video with eccentricity will introduce aliasing and jitter effect when viewed. Guenter *et al.* [15] progressively compute three gaze-centered concentric rings to address this problem. Stengel *et al.* [40] propose to perform sparse rendering in the periphery with either stochastic sampling and inpainting. Temporal models from VSR are referred for the design of models for FVSR [5, 47].

Single Image Super-Resolution. Early work on super-resolution processes each frame separately. SRCNN is a simple 3-layer super-resolution convolutional neural network proposed by Dong *et al.* [10]. Kim *et al.* [27] explores a deeper architecture, VDSR, a 20-layer deep network with residual connections. ResNet [18] and generative adversarial networks [14] is adopted by Ledig *et al.* [29] in SRGAN and Sajjadi *et al.* [38] in EnhanceNet to generate high-frequency detail. Tai *et al.* [42] propose DRRN that uses recursive residual blocks. Tong *et al.* propose SR-

DenseNet [45] which uses DenseNet [19] as its backbone. Pan *et al.* [33] propose DualCNN that uses two branches to reconstruct structure and detail components of an image.

Video Super-Resolution. To exploit temporal information across frames in a video, temporal alignment or motion compensation is used either explicitly or implicitly. Explicit VSR makes use of information from neighboring frames through motion estimation and compensation. Earlier work for motion estimation is based on optical flow, e.g. Liao *et al.* [31] uses optical flow methods [52] along with a deep draft-ensemble network to reconstruct the HR frame. Kappler *et al.* [25] predicts HR frame by taking interpolated flow-wrapped frames as inputs to a CNN. VESPCN [3] is the first VSR work that jointly trains flow estimation and spatio-temporal networks. SPMC [43] uses an optical flow network to compute LR motion field to generate sub-pixel information to achieve sub-pixel motion compensation. TOFlow [53] shows that task-oriented motion cues achieves better VSR results than fixed flow algorithms. RBPN [16] propose a recurrent encoder-decoder module to exploit inter-frame motion that is estimated explicitly. EDVR [47] uses a deformable convolution module to align multiple frames to a reference frame in feature space and uses a temporal and spatial attention module for fusion. Jo *et al.* [23] are the first to propose the use of dynamic up-sampling filters (DUF) for VSR. To consider neighboring frames for implicit VSR, a common approach is by using a sliding window, i.e the concatenation of images within a fixed window length [53, 23, 16, 47, 21]. Most sliding window methods are symmetric, i.e. past and future frames are considered for the reconstruction of the targeted frame (non-causal), making them unsuitable for streaming applications. Recurrent methods [39, 13, 20] passes information from previous frames through a hidden representation. Our work is based on the idea of flow-guided deformable alignment from BasicVSR++ [5]. Such idea has been earlier studied by several works [22, 41].

3. Methods

We propose Cross-Resolution Flow Propagation (CRFP), a novel framework for FVSR. CRFP is able to aggregate context from gaze region that is of high resolution (HR) to the low resolution (LR) counterpart. To provide high fidelity video stream, HR context of previous frames should be captured and retained by the framework such that future frames can be better super-resolved using the retained context. This design works in tandem with the nature of eye tracking devices. The gaze coordinate predicted by eye tracking devices is usually corrupted by additive Gaussian noise, $\mathbf{p}_t^{\text{Fov}} \in \mathcal{N}(\mu_t^{\text{Fov}}, \sigma_t^{\text{T}})$, having the predicted gaze coordinate oscillating around the actual gaze direction μ_t^{Fov} under a Gaussian noise of the eye

tracker of standard deviation σ_t^{T} . This is analogous to the application of super-resolution techniques to handheld cameras [50, 28, 1], where the natural hand tremor is exploited during the reconstruction of the original frame. The better a model is at capturing and retaining HR context from past foveated region, the better it can exploit the prediction noise from the eye tracker. In Section 3.1 we discuss the pathways in our architecture that contributes to the retention of context from past foveated region. In Section 3.2 we provide in-depth description of the Feature Aggregator. In Section 3.3 we show how context from the foveated region is aggregated into the feedback and frame generation pipeline. We illustrate an overview of CRFP in Figure 1.

3.1. Cross-Resolution Flow Propagation

To adopt the HR context for the super-resolution of LR context corresponding to current and future frames, an architecture that focuses on cross-resolution propagation of context is required. CRFP is proposed to handle this problem, introducing two core building blocks for cross-resolution context aggregation, namely the *Feature Aggregator (FA)* and *Output Block (OB)*. These building blocks are connected by several information pathways, i.e. *Feedback Pathway*, *DCN Propagation Pathway*, *Flow Field Pathway*, *Warped Feature Pathway*, *Fovea Pathway* and *Low-Res Feature Pathway*, each playing different roles in the aggregation process. The goal of FVSR is to super-resolve the LR frame at timestep t , \mathbf{x}_t^{LR} , while considering an additional foveated region of HR, $\mathbf{x}_t^{\text{Fov}}$. Without any external factors, the LR frame \mathbf{x}_t^{LR} propagates through the Low-Res Feature Pathway. \mathbf{x}_t^{LR} is first encoded by an encoder \mathcal{E}^{LR} followed by a pixel shuffle + convolution block $\mathcal{S}_{\text{s}\uparrow}^2$ for $2\times$ up-sampling, giving us \mathbf{h}_t^0 ,

$$\mathbf{h}_t^0 = \mathcal{S}_{\text{s}\uparrow}^2(\mathcal{E}^{\text{LR}}(\mathbf{x}_t^{\text{LR}})). \quad (1)$$

The encoded features are then fed to several FA blocks along the Low-Res Feature Pathway to be aggregated with information from other pathways,

$$\begin{aligned} \{\mathbf{h}_t^{l+1}, \mathcal{D}_t^{l+1}\} = \\ \text{FA}^l(\mathcal{S}_{\text{s}\downarrow}^4(\mathbf{h}_t^l); \hat{\mathbf{h}}_{t-1}, \mathbf{F}_t, \mathcal{S}_{\text{s}\downarrow}^4(\tilde{\mathbf{h}}_{t-1}), \mathcal{D}_t^l, \mathcal{W}(\mathbf{z}_t^l; \mathbf{F}_t)), \\ l = 0, \dots, L-1. \end{aligned} \quad (2)$$

In our design, we have $L = 4$ where the first three FAs are placed at the feature of the lowest spatial resolution and one FA placed after the up-sampling stage $\mathcal{S}_{\text{s}\uparrow}^4$. The motivation of such placement is to enable the aggregation of information at different spatial resolution while keeping computational cost low. Placing FA after the up-sampling stage would increase the computational cost quadratically in accordance to the up-sampling rate but the aggregation of HR

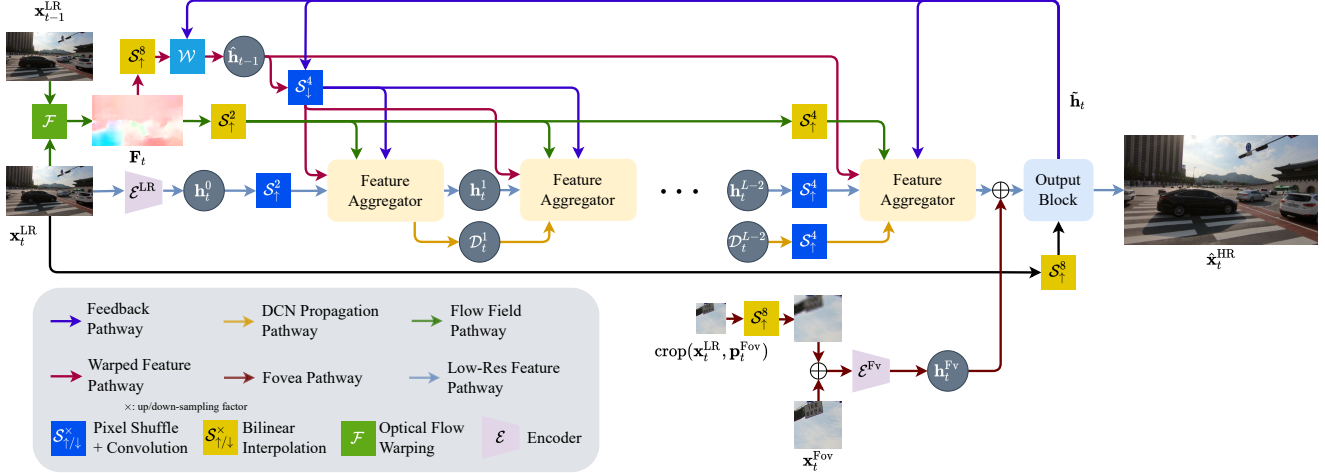


Figure 1. Overview of Cross-Resolution Flow Propagation for $8\times$ FVSR. The core building blocks are **Feature Aggregator (FA)** and **Output Block (OB)**. Foveated region is regionally aggregated at the **OB**. The **OB** has two outputs, the super-resolved frame $\hat{\mathbf{x}}_t^{\text{HR}}$ and the fovea-aggregated feature $\tilde{\mathbf{h}}_t$ that is fed to earlier stages of the network through the **Feedback Pathway**. Features at the lower spatial resolution encoded by \mathcal{E}^{LR} and features of the highest spatial resolution right after the upsampling (**Pixel Shuffle + Convolution**) block $\mathcal{S}_{s\downarrow}^4$ that passes through the **Low-Res Feature Pathway**, are aggregated with features from other pathways (**Feedback**, **DCN Propagation**, **Flow Field**, **Warped Feature**) through the **FA**. Repeated adoption of **FA** sharing the same input and output connection pattern is abbreviated as “...” in the illustration. We use the same downsampling (**Pixel Shuffle + Convolution**) block $\mathcal{S}_{s\downarrow}^4$ (tied-weights) to encode features from the **Feedback Pathway** and **Warped Feature Pathway** for the **FA**.

context is more precise since it’s closer to the coordinate system of the pixel space. The output of the final FA \mathbf{h}_t^{L-1} is concatenated (\oplus) with the encoded foveated region \mathbf{h}_t^{Fv} as input for the OB to render the super-resolved frame $\hat{\mathbf{x}}_t^{\text{HR}}$ and to estimate the fovea-fused feature $\tilde{\mathbf{h}}_t$ to be propagated to earlier layers through the Feedback Pathway for aggregation with features of future frames,

$$\{\hat{\mathbf{x}}_t^{\text{HR}}, \tilde{\mathbf{h}}_t\} = \text{OB}(\mathbf{h}_t^{L-1} \oplus \mathbf{h}_t^{\text{Fv}}, \mathcal{S}_{s\uparrow}^8(\mathbf{x}_t^{\text{LR}})). \quad (3)$$

Note that OB also takes in the bilinearly-upsampled LR frame $\mathcal{S}_{s\uparrow}^8(\mathbf{x}_t)^{\text{LR}}$ with details deferred to Section 3.3. \mathbf{h}_t^{Fv} originates from the Fovea Branch,

$$\mathbf{h}_t^{\text{Fv}} = \mathcal{E}^{\text{Fv}}(\mathbf{x}_t^{\text{Fv}} \oplus \text{crop}(\mathcal{S}_{s\uparrow}^8(\mathbf{x}_t^{\text{LR}}), \mathbf{p}_t^{\text{Fv}})). \quad (4)$$

\mathbf{h}_t^{Fv} is the result of a fovea encoder \mathcal{E}^{Fv} applied to the concatenation of the HR foveated region \mathbf{x}_t^{Fv} and the $8\times$ bilinearly up-sampled LR frame \mathbf{x}_t^{LR} cropped (crop) using gaze coordinate \mathbf{p}_t^{Fv} . The fovea-fused feature $\tilde{\mathbf{h}}_{t-1}$ from the previous time-step is down-sampled using a pixel shuffle + convolution block $\mathcal{S}_{s\downarrow}^4$ to match the spatial resolution of the features of earlier layers in the FA. The same downsampling block is utilized to down-sample the warped version of \mathbf{h}_{t-1} ,

$$\hat{\mathbf{h}}_{t-1} = \mathcal{W}(\tilde{\mathbf{h}}_{t-1}, \mathcal{S}_{s\uparrow}^8(\mathbf{F}_t)). \quad (5)$$

\mathcal{W} is the warping operator that warps an input image using optical flow \mathbf{F}_t . \mathbf{F}_t is bilinearly up-sampled using $\mathcal{S}_{s\uparrow}^8(\cdot)$ to

match the size of $\tilde{\mathbf{h}}_{t-1}$. \mathbf{F}_t is estimated using an optical flow estimator \mathcal{F} based on frames from time-steps t and $t-1$,

$$\mathbf{F}_t = \mathcal{F}(\mathbf{x}_t^{\text{LR}}, \mathbf{x}_{t-1}^{\text{LR}}). \quad (6)$$

The flow field \mathbf{F}_t is also bilinearly up-sampled $\mathcal{S}_{s\uparrow}^2$ to match the spatial resolution of features in the FAs. As there is a Deformable Convolutional Layer (DCN) embedded within the FA, we pass the feature responsible for the generation of DCN parameters, i.e offsets and masks, across FAs through the DCN Propagation Pathway, acting as residual connection, also known as *Flow Propagation*,

$$\{\mathbf{h}_t^{l+1}, \mathcal{D}_t^{l+1}\} = \text{FA}^l(\mathcal{S}_{s\downarrow}^4(\mathbf{h}_t^l); \cdot, \cdot, \cdot, \mathcal{D}_t^l, \cdot), \quad l = 0, \dots, L-1, \quad (7)$$

$$\mathcal{D}_t^0 = \mathbf{0}. \quad (8)$$

All FAs except the FA after the up-sampling stage shares the same input configuration. For the final FA, the Low-Res Feature \mathbf{h}_t^{L-2} and DCN parameters \mathcal{D}_t^{L-2} are independently $4\times$ up-sampled using pixel shuffle and convolution. The previously $2\times$ bilinearly up-sampled flow field is further $4\times$ bilinearly up-sampled. The fovea-fused feature $\tilde{\mathbf{h}}_{t-1}$ and $\hat{\mathbf{h}}_{t-1}$ bypasses the down-sampler $\mathcal{S}_{s\downarrow}^4$ in the earlier layers and are fed directly into the final FA.

3.2. Feature Aggregator

With the high-level connections between modules defined, we discuss the inner workings of FA here. An illustration of FA is shown in Figure 2. In FA, DCN state

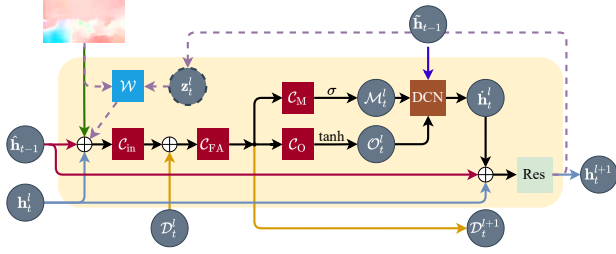


Figure 2. Illustration of **Feature Aggregator (FA)**. C 's are convolutional layers, DCN is the Deformable Convolutional Layer and Res is the residual block. DCN warps the fovea-aggregated feature \tilde{h}_{t-1} from the **Feedback Pathway** using the estimated offsets \mathcal{O}_t^l and is weighted by estimated masks \mathcal{M}_t^l . The result of DCN is concatenated with features from the **Warped Feature Pathway** \hat{h}_{t-1} and the **Low-Res Feature Pathway** h_t^l and fed to a Residual Block to predict the Low-Res Feature of the upcoming stage h_t^{l+1} and the **optional DCN state vector (DSV)** z_t^{l+1} . Each FA has its own DSV that involves in the estimation of masks and offsets of DCN. Variables and connections that involves the DSV are represented with **dashed outline/line**.

vector (DSV) z_t^l along with features from the Warped Feature Pathway \hat{h}_{t-1} , Low-Res Feature Pathway h_t^l and Flow Field Pathway \mathbf{F}_t are responsible for the estimation of masks \mathcal{M}_t^l and offsets \mathcal{O}_t^l required for DCN,

$$\mathcal{D}_t^{l+1} = C_{FA}^l \left(C_{in}^l \left(\hat{h}_{t-1} \oplus h_t^l \oplus \mathcal{S}_{1f}^u(\mathbf{F}_t) \oplus \mathcal{W}(z_t^l; \mathcal{S}_{1f}^u(\mathbf{F}_t)) \right) \oplus \mathcal{D}_t^l \right), \quad (9)$$

$$\mathcal{M}_t^l = \sigma \left(C_M^l(\mathcal{D}_t^{l+1}) \right), \quad (10)$$

$$\mathcal{O}_t^l = \tanh \left(C_O^l(\mathcal{D}_t^{l+1}) \right). \quad (11)$$

Note that $u = 8$ if $l = L - 2$ and $u = 2$ otherwise, C^l are convolutional blocks and \mathcal{D}_t^l is the feature responsible for the estimation of \mathcal{M}_t^l and \mathcal{O}_t^l . \mathcal{D}_t^l is passed to the upcoming DCN block as residual connection. We can then perform DCN on the feature from the Feedback Pathway \tilde{h}_{t-1} (down-sampled with $\mathcal{S}_{s\downarrow}^4$ for $l < L - 2$) as follows,

$$\hat{h}_t^l = \text{DCN}(\tilde{h}_{t-1}; \mathcal{M}_t^l, \mathcal{O}_t^l), \quad (12)$$

$$= C_{DCN}^l \left(\mathcal{M}_t^l \odot \mathcal{W}(\tilde{h}_{t-1}; \mathcal{O}_t^l) \right), \quad (13)$$

$$\hat{h}_t^l(p) = \sum_{k=1}^K \mathbf{w}_k^l \cdot \tilde{h}_{t-1}(p + p_k + \mathcal{O}_t^l(p)) \cdot \mathcal{M}_t^l(p). \quad (14)$$

Finally, we can estimate DSV for the next time-step z_{t+1}^l along with the low-res feature for the upcoming stage h_t^{l+1} by passing the DCN-warped feature \hat{h}_t^l along with features from the Warped Feature Pathway \hat{h}_{t-1} and Low-Res Feature Pathway h_t^l to a Residual Block,

$$\{h_t^{l+1}, z_{t+1}^l\} = \text{Res} \left(\hat{h}_t^l \oplus \hat{h}_{t-1} \oplus h_t^l \right). \quad (15)$$

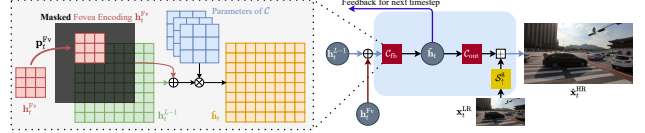


Figure 3. Illustration of **Output Block (OB)**. Given the gaze coordinate \mathbf{p}_t^{Fv} , the encoded foveated region features h_t^{Fv} are aligned to the designated position where regional convolution is applied on the concatenated features to generate the fovea-fused feature \tilde{h}_t . \tilde{h}_t is passed through the **Feedback Pathway** to provide context from the foveated region to the **FA** blocks. Finally, the frame is rendered at the targeted resolution by adding the estimated result as residual to the bilinearly up-sampled LR frame $\mathcal{S}_{1f}^8(x_t^{LR})$.

DSV is helpful in retaining information that might potentially be corrupted by the warping operation of the upcoming time-step, e.g. future occlusion.

3.3. Output Block

After passing through several stages of FAs, comes the final stage where the frame at the targeted resolution will be rendered using the Output Block (OB). We show an illustration of OB in Figure 3. The OB takes in h_{t-1}^{L-1} from the Low-Res Feature Pathway and h_t^{Fv} from the Fovea Pathway along with the bilinearly up-sampled LR frame $\mathcal{S}_{1f}^8(x_t^{LR})$ for frame rendering,

$$\tilde{h}_t = C_{fb} \left(h_{t-1}^{L-1} \oplus h_t^{Fv} \right), \quad (16)$$

$$\hat{x}_t^{HR} = C_{out} \left(\tilde{h}_t \right) + \mathcal{S}_{1f}^8 \left(x_t^{LR} \right). \quad (17)$$

The estimated results acts as residuals that enhance the bilinearly up-sampled frame. \tilde{h}_t contains context from the foveated region and is responsible for the enhancement of feature from earlier stages of the upcoming time-steps. The quality of \tilde{h}_t affects the capability of a model to retain HR context that will potentially be picked-up and utilized by the FAs of earlier stages. As prediction of gaze coordinates using eye trackers usually comes with noise, the capability of a model to retain past HR context is beneficial for the task of FVSR, affecting the overall visual fidelity of transmitted foveated video stream.

4. Experiments

This work studies the novel task of FVSR. Experimental setup is referred from the task of VSR, with several tweaks to shift the focus of experiments towards FVSR. Since FVSR is targeted for video streaming applications on AR/VR devices that are paired with an eye tracker, we design our experiments to fit such use case through the demonstration of the retention of past foveated region.

Dataset. Our experiment setup for FVSR bootstraps on the VSR experiments of BasicVSR [4, 5]. We train and



Figure 4. Visual comparison on the task of $8\times$ FVSR. The foveated region (red box) slides from left to right from frame t to frame $t + T$ where $T = 6$. PSNR and SSIM plots are for performance comparison, with brighter pixel value indicating higher performance.

evaluate on REDS [32]. We use REDS4¹ as our test set and REDSval4² as our validation set. The remaining clips are used as our training set. We carry out our study on $8\times$ FVSR since less visual attention is allocated to the region beyond the foveated region. We apply $8\times$ Bicubic Down-sampling to obtain the LR frame for FVSR.

Architecture. We use an encoder-decoder for the optical flow estimator \mathcal{F} that is separately trained on MPI Sintel dataset [2]. Each convolution block is composed of a single convolutional layer paired with leaky ReLU as its non-linear activation function. Each up/down-sampling block is composed of a single convolutional layer paired with pixel shuffle operation. The encoders \mathcal{E}^{LR} and \mathcal{E}^{Fv} are composed of 2 convolutional blocks. For comparison purpose, we modify BasicVSR++ [5] for the task of FVSR. BasicVSR++ is made causal and foveated region is fed directly to the layer of lowest spatial resolution. For computational efficiency, we allocate three FA blocks to the up-sampled encoded LR frame and a single FA block right before the output block. Detailed configuration is deferred to the supplementary materials.

Training and evaluation. We train our models using PyTorch [34] as our deep learning framework using a single RTX3090 GPU. Runtime in 1 are measured using frame of size 1080p. The initial learning rate of our model and flow estimator are set to 1×10^{-4} and 2.5×10^{-5} respectively.

¹Clips 000, 011, 015, 020 of REDS training set.

²Clips 000, 001, 006, 017 of REDS validation set.

The total number of training iterations is 300K with a batch size of 8. We use Charbonnier loss [6] as our loss function for better robustness against outliers. Images of RED4 are of size 1280×720 . We crop regions of size 256×256 which are down-sampled to 32×32 as our LR frame during training. We also crop regions of size 128×128 from the 256×256 patch to be our foveated region. For training, the coordinate of the foveated region is randomly sampled across whole image. The coordinate of the foveated region is constrained to not move out of boundary. For evaluation, we show results that slides foveated region in a raster scan order. We also show results that has the fovea coordinate oscillate in the vicinity of additive Gaussian noise present in eye trackers to demonstrate the actual use case of CRFP for FVSR task. The foveated region is cropped to represent $\sim 1\%$ of the total pixels in a HR image. We down-sample the LR frames to be of size 160×90 and the cropped foveated region size is 96×96 . We use PSNR, SSIM and VMAF [30] as metrics to measure the quality of the super-resolved frame. VMAF is targeted towards video stream and is shown to have higher correlation to the human perception of visual fidelity of video when compared to PSNR and SSIM.

4.1. In-Depth Study of Foveated Video Super-Resolution

To evaluate the effectiveness of FVSR, we need to measure the capability of method in retaining HR context of the foveated region and propagating it to future frames. The fusion of both HR and LR context is also important for op-

timal visual fidelity. We can evaluate FVSR using these two regions:

1. Foveated region: measures the efficiency in the fusion and the transferring of HR context to the current LR frame.
2. Past foveated region(s): measures the efficiency in the retaining of HR context of past foveated frame(s) and propagating it to future frames.

As there is no prior work in this field, a fair comparison would be adopting a modification of the SoTA in VSR, BasicVSR++ [5], to fit the task of FVSR. We name the modified version as BasicFVSR++. We modify it to be causal, i.e. only frames prior to the current frames are considered for FVSR. We also reduce its size for the ease of experimentation. Foveated region is fused with features of earlier layers which differs from our contribution that considers the fusion on a higher spatial resolution. There are several variations of CRFP that are ablated and studied. CRFP (removal of flow propagation) corresponds to the removal of the DCN Propagation Pathway. CRFP correspond to the vanilla version that doesn't include the DSV. CRFP + DSV (no fovea) corresponds to the study of CRFP that includes DSV but without the Fovea Pathway. CRFP + DSV correspond to the inclusion of the optional DSV. CRFP-Fast includes DSV and applies the DCN blocks within FA only to a fixed region of size 720×720 to focus on low-latency FVSR. Quantitative analysis of all regions are summarized in Table 1. From the results, we can see that CRFP is better than BasicFVSR++. We show the runtime and parameter count of different models for FVSR in Table 2. In Figure 4, we also show a qualitative comparison of all methods on two different time-steps with a frame interval of $T = 6$ in between. CRFP is visually better when compared to BasicFVSR++. From the SSIM plots, we can clearly observe past HR contexts are better retained.

Foveated Region. We do not directly place the HR frame (foveated region) onto the super-resolved frame to prevent sharp transition of quality across regions of different resolutions. With the HR frame propagated through a series of convolutional layers, there will definitely be a slight drop in quality as convolutional filtering is a noisy process. To measure the efficiency in the propagation of fovea information in the main branch, we use PSNR and SSIM as metrics to evaluate the pixels that fall in the foveated region. Our results show that CRFP outperforms BasicFVSR++ by 4.02 dB. DSV module is also beneficial for the propagation of fovea information across the network with less parameters, showing a marginal boosts of 0.13 dB. Qualitative results in Figure 4 show that finer details can be reconstructed by CRFP when compared with BasicFVSR++.



Figure 5. Comparison visual quality of past foveated regions across different methods. Two regions containing high frequency contents are shown. Refer to the license plate of the car and the tiles on the pavement to spot the distinction across different methods.

Past Foveated Regions. An important property of FVSR is the capability of a model to retain HR context(s) that corresponds to foveated region of previous time-step(s). To measure this property, we slide the foveated region across frames using a horizontal trajectory, i.e. a straight line from left to right. The region covered by the trajectory of the foveated region should retain past HR context with high probability. We evaluate the retention capability of a FVSR model by measuring the PSNR and SSIM at the region covered by the fovea trajectory. Our results show that CRFP is much better than BasicFVSR++ at retaining information from previous frames, showing 0.25 dB increase in PSNR. The inclusion of the DSV module has similar performance as the vanilla variant while requiring less parameters. The effect of past frame retention is more prominent in the visualizations shown in Figure 5. We can see that fine-details from previous HR contexts are still present after an interval

Table 1. Performance comparison of $8\times$ FVSR evaluated using REDS4 at proposed regions using PSNR, SSIM and VMAF.

| Method | Foveated Region | | Past Foveated Region(s) | | Whole Image | | |
|-----------------------|-----------------|---------------|-------------------------|---------------|--------------|---------------|--------------|
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | VMAF |
| Bicubic | 26.16 | 0.6077 | 24.72 | 0.5994 | 23.34 | 0.6077 | 5.1621 |
| BasicFVSR++ [5] | 37.99 | 0.9560 | 30.36 | 0.8269 | 25.95 | 0.7250 | 68.40 |
| CRFP + DSV (no fovea) | 29.23 | 0.7128 | 28.24 | 0.7187 | 25.52 | 0.7001 | 64.38 |
| CRFP | 42.07 | 0.9831 | 30.22 | 0.8337 | 25.84 | 0.7202 | 66.32 |
| CRFP | 42.01 | 0.9835 | 30.61 | 0.8451 | 26.13 | 0.7336 | 70.24 |
| CRFP-Fast | 42.14 | 0.9831 | 29.44 | 0.7983 | 23.72 | 0.6365 | 24.40 |
| CRFP + DSV | 42.14 | 0.9836 | 30.59 | 0.8455 | 26.07 | 0.7338 | 70.30 |

Table 2. Runtime and model parameters comparison for 1080p video using Nvidia RTX 3090.

| Method | Runtime (ms) | # Parameters |
|-----------------------|--------------|--------------|
| BasicFVSR++ [5] | 35 | 2.35M |
| CRFP + DSV (no fovea) | 41 | 2.17M |
| CRFP | 39 | 2.16M |
| CRFP | 42 | 2.21M |
| CRFP-Fast | 14 | 2.17M |
| CRFP + DSV | 41 | 2.17M |

of a $T = 9$ frames.

4.2. Simulating FVSR with Eye Tracker Noise

To simulate the application of CRFP to an actual use case of FVSR, we follow the pattern found in eye trackers to influence the trajectory of the foveated region’s coordinates. In Figure 6, we show results for coordinates oscillating under an additive Gaussian noise of $\sigma^T = 10$, $\sigma^T = 50$ and $\sigma^T = 100$. We can observe that with $\sigma^T = 100$ a larger region can be super-resolved with the context transferred from the past foveated regions. This experiment shows that the capability of a model on retaining context from past foveated region is a good measure of performance and transfers well to the task of FVSR. The design of this experiment is to demonstrate the importance of the transferring of context from past foveated region to future frames for the task of FVSR. The better the performance in retaining context from previous frames the more resilient it is to the noise present in trackers. The results with $\sigma^T = 100$ demonstrates that context surrounding the gaze region can be clearly reconstructed despite high variation in the eye tracker’s reading.

5. Conclusion

We propose a novel research direction of Foveated Video Super-Resolution (FVSR) with reliable metrics for the measurement of the foveated visual quality. The measurement of quality of past foveated region is shown to be beneficial for the task of FVSR through experiments that simulates eye

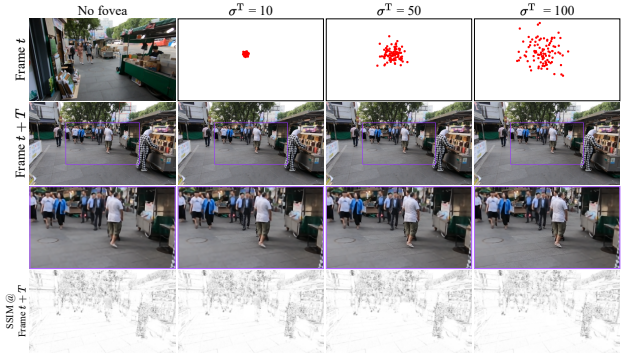


Figure 6. Simulating the actual use case of FVSR where there is additive Gaussian noise present in an eye tracker. Various standard deviations σ^T are tested and results show that larger σ^T demonstrates the capability of the model on retaining HR context from past foveated regions. Spot the difference on the stripes on the man’s shirt and the lines on the ground. SSIM plots are also provided to assist the reader in spotting the differences across different σ^T .

tracker noise. We show that with induced additive Gaussian noise, CRFP is able to super-resolve context that falls within the specified region through the adoption of context from previous HR foveated region. Under these metrics, we demonstrate that CRFP is able to perform well on the task of FVSR. CRFP is designed specifically for FVSR and is suitable for video streaming in AR/VR applications. CRFP is also designed to be of low-latency, suitable for head-mounted AR/VR devices with limited computational capacity.

Acknowledgement

This project is supported by MOST under code MOST 110-2221-E-A49-144-MY3. Eugene Lee is partially supported by Novatek Ph.D. Fellowship Award. The authors are grateful for the suggestions provided by Dr. Eugene Wong from University of California in Berkeley and Dr. Jian-Ming Ho from Academia Sinica of Taiwan.

A. 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}^{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}^{Fv} have output channel of size 4. \mathcal{C}_{fb} has input channel of 8 (concatenation of foveated region features and features from feature aggregator) and output channel of 4. \mathcal{C}_{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^{LR} and $\mathbf{I}_{t-1}^{\text{LR}}$, 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.

B. 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 3. We can observe that the introduction of a small amount of DSV into the feature aggregator contributes to the final performance.

C. Simulating FVSR with Eye Tracker Noise

We show similar a simulation as the one shown in the main paper in Figure 7.

D. 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 8. 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.

E. Video in Supplementary Materials

We show videos with the format of Figure 8 in our supplementary materials. The name formatting of the videos follows the rule $\sigma^T_{\text{videoID}}.\text{mp4}$. σ^T correspond to the standard deviation of the distribution of the additive Gaussian noise introduced to the foveated region.

References

- [1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021.
- [2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012.
- [3] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4778–4787, 2017.
- [4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021.
- [5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. *arXiv preprint arXiv:2104.13371*, 2021.
- [6] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 168–172. IEEE, 1994.
- [7] Viviane Clay, Peter König, and Sabine Koenig. Eye tracking in virtual reality. *Journal of eye movement research*, 12(1), 2019.
- [8] Christine A Curcio, Kenneth R Sloan, Robert E Kalina, and Anita E Hendrickson. Human photoreceptor topography. *Journal of comparative neurology*, 292(4):497–523, 1990.
- [9] Dennis M Dacey and Michael R Petersen. Dendritic field size and morphology of midget and parasol ganglion cells of the human retina. *Proceedings of the National Academy of sciences*, 89(20):9666–9670, 1992.

Table 3. 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 |

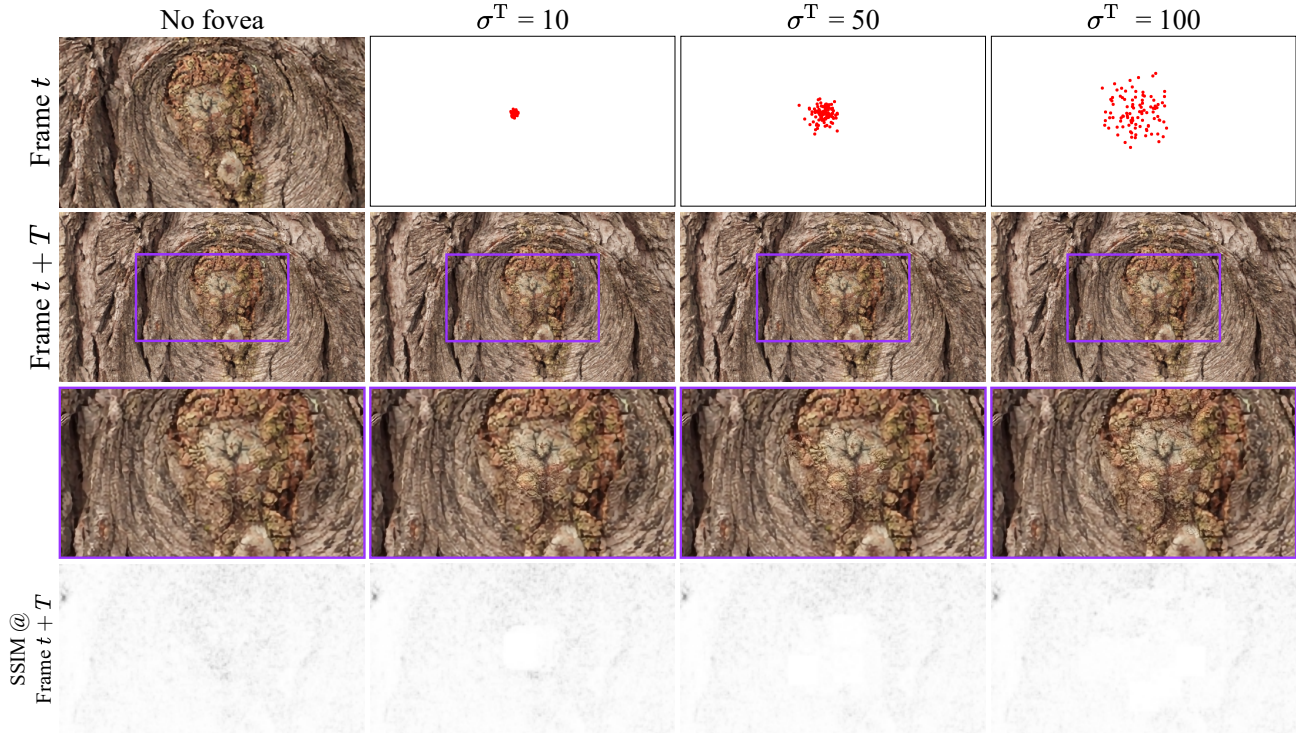


Figure 7. Simulating the actual use case of FVSR where there is additive Gaussian noise present in an eye tracker. Various standard deviations σ^T are tested and results show that larger σ^T demonstrates the capability of the model on retaining HR context from past foveated regions. Spot the difference 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.

- [10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.
- [11] Andrew Edelsten, Paula Jukarainen, and Anjul Patney. Truly next-gen: Adding deep learning to games and graphics. In *NVIDIA Sponsored Sessions (Game Developers Conference)*, 2019.
- [12] Manuel Fernandez. Augmented virtual reality: How to improve education systems. *Higher Learning Research Communications*, 7(1):1–15, 2017.
- [13] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019.
- [14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.
- [15] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. Foveated 3d graphics. *ACM Transactions on Graphics (TOG)*, 31(6):1–10, 2012.
- [16] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019.
- [17] Azizul Hassan, Erdogan Ekiz, Sumesh S Dadwal, and Geoff Lancaster. Augmented reality adoption by tourism product and service consumers: Some empirical findings. In *Augmented Reality and Virtual Reality*, pages 47–64. Springer, 2018.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

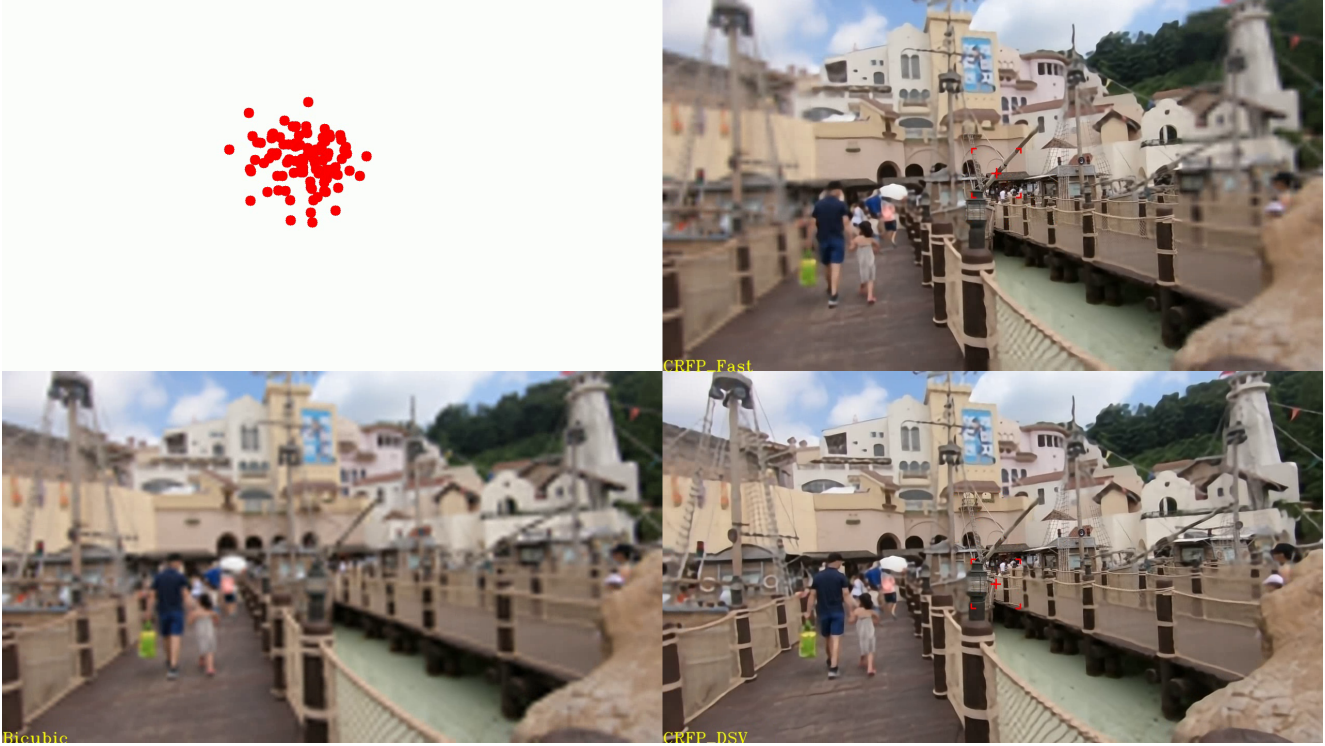


Figure 8. **Top Left:** 100 center points of foveated region; **Top Right:** CRFP-Fast after 100 frames; **Bottom Right:** CRFP + DSV after 100 frames; **Bottom Left:** Bicubic result after 100 frames. Notice that while CRFP has noticeably lower quality beyond the region (720×720) passed into the DCN, it is not visually perceptible if we focus on the foveated region.

ings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

- [19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [20] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *European Conference on Computer Vision*, pages 645–660. Springer, 2020.
- [21] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8008–8017, 2020.
- [22] Zhiwei Jia, Haoshen Hong, Siyang Wang, Kwonjoon Lee, and Zhuowen Tu. Controllable top-down feature transformer. *arXiv preprint arXiv:1712.02400*, 2017.
- [23] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3232, 2018.
- [24] Anton S Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. Deep-fovea: neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.
- [25] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016.
- [26] DH Kelly. Retinal inhomogeneity. i. spatiotemporal contrast sensitivity. *JOSA A*, 1(1):107–113, 1984.
- [27] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.
- [28] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2370–2379, 2021.
- [29] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.
- [30] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2), 2016.

- [31] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 531–539, 2015.
- [32] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [33] Jinshan Pan, Sifei Liu, Deqing Sun, Jiawei Zhang, Yang Liu, Jimmy Ren, Zechao Li, Jinhui Tang, Huchuan Lu, Yu-Wing Tai, et al. Learning dual convolutional neural networks for low-level vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3070–3079, 2018.
- [34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [35] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016.
- [36] John G Robson. Spatial and temporal contrast-sensitivity functions of the visual system. *Josa*, 56(8):1141–1142, 1966.
- [37] Jyrki Rovamo, Lea Leinonen, Pentti Laurinen, and Veijo Virsu. Temporal integration and contrast sensitivity in foveal and peripheral vision. *Perception*, 13(6):665–674, 1984.
- [38] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.
- [39] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6626–6634, 2018.
- [40] Michael Stengel, Steve Grogoric, Martin Eisemann, and Marcus Magnor. Adaptive image-space sampling for gaze-contingent real-time rendering. In *Computer Graphics Forum*, volume 35, pages 129–139. Wiley Online Library, 2016.
- [41] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.
- [42] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.
- [43] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4472–4480, 2017.
- [44] Larry N Thibos, David L Still, and Arthur Bradley. Characterization of spatial aliasing and contrast sensitivity in peripheral vision. *Vision research*, 36(2):249–258, 1996.
- [45] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE international conference on computer vision*, pages 4799–4807, 2017.
- [46] Lingdong Wang, Mohammad Hajiesmaili, and Ramesh K Sitaraman. Focas: Practical video super resolution using foveated rendering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5454–5462, 2021.
- [47] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [48] Yi-Zhong Wang, Larry N Thibos, and Arthur Bradley. Undersampling produces non-veridical motion perception, but not necessarily motion reversal, in peripheral vision. *Vision Research*, 36(12):1737–1744, 1996.
- [49] Wei Wei. Research progress on virtual reality (vr) and augmented reality (ar) in tourism and hospitality: A critical review of publications from 2000 to 2018. *Journal of Hospitality and Tourism Technology*, 2019.
- [50] Bartłomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019.
- [51] Lei Xiao, Salah Nouri, Matt Chapman, Alexander Fix, Douglas Lanman, and Anton Kaplanyan. Neural supersampling for real-time rendering. *ACM Transactions on Graphics (TOG)*, 39(4):142–1, 2020.
- [52] Li Xu, Jiaya Jia, and Yasuyuki Matsushita. Motion detail preserving optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1744–1757, 2011.
- [53] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.