Human Vision Constrained Super-Resolution

Supplementary Material

Abstract

This supplementary file presents further details and additional results of the proposed method. These results illustrate further outcomes of the user study. Subsequently, the preliminary prototype of the method applied to AR/VR headsets are illustrated. Finally, further qualitative results of the perceptual model are presented to demonstrate the efficacy of the method.

A. Subjective Quality Study Setup

Setup. For our experiments, we assumed standard office conditions where the content is viewed on a 27-inch Dell U2723QE display with a resolution of 3840 $\times$ 2160 and a peak luminance of 400 cd/m^2, from a viewing distance of 60 cm. All our results are calculated according to this setup. During our experiments with human subjects, the viewing distance was controlled with the use of a chin rest that allowed to maintain constant viewing conditions throughout the experiments for all participants.

B. Additional Subjective Quality Results

B.1. Channel Depth Application

For this experiment, we used the same scenes as described in Sec. 5. We up-scaled the images using a set of EDSR networks. We trained 5 different networks, with either (256, 128, 64, 16, 8) channels per-layer. The baseline was an EDSR with 256 channels per layer applied uniformly across the whole image. The baseline was compared with SR controlled using our perceptual model, which selected one of the 5 candidate networks per each patch. Fig. 1 shows the results of our 2AFC user study, it can be seen that on average, the preference value hovers around 50%, which is indicative that users wone average not able to perceive any difference between the test cases A and B in relation to the reference. The indicates that our perceptual model controlled the results such that any quality loss is not visible, while using 76.4% less FLOPS.

B.2. Video Content

Setup and Task The same experimental protocol as described in Sec. 5 was employed for this experiment.
Stimuli The stimuli were derived from seven natural videos from the Inter4k [4] dataset.
Data preparation Each video frame was down-sampled by a factor of eight and then up-sampled by a factor of four using the VDSR network, specifically the network branching application.

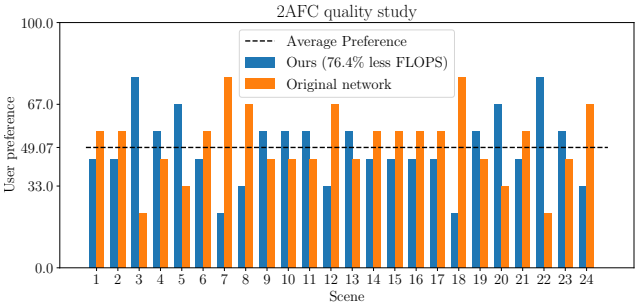

Figure 1. The result of our subjective study (for 9 participants) for the network channel depth application.

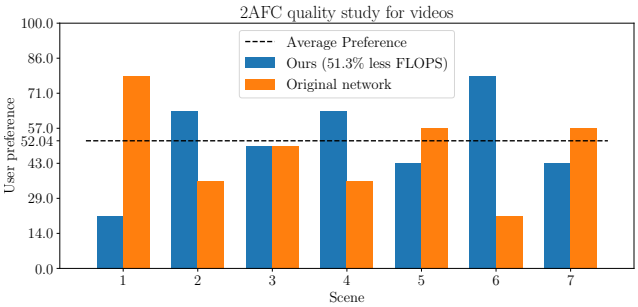

Figure 2. The result of our subjective study (for 14 participants) for the network branching application for videos.

Fig. 2 shows the findings of our 2AFC user study. It can be observed that the mean preference is around 50%, which suggests that users were unable to perceive a difference between test cases A and B in relation to the reference, even when our method uses 51.3% less FLOPS.

B.3. Information about participants

All participants were recruited from CS department, aged 20 to 30. The network branching user study was conducted with a total of 15 participants (4F, 11M). The channel depth user study was conducted with a sample of 9 participants (3F, 6M)). Finally, the network branching applied to videos user study was conducted with 14 participants (3F, 11M). They had normal or corrected-to-normal vision and were unaware of the experiment's purpose.

C. FLOPS for evaluation and performance bottleneck

We decided to use the average FLOPS for efficiency due to its universality, independence from specific machine characteristics and implementation, as well as the fact that this measure has been used in similar studies [2, 3, 5, 7]. In our method patches that require the use of a network with larger size/capacity could be considered as a bottleneck, and in such a case, using the average FLOPS does not convey this information. A wall clock time elapsed could be considered a better measure of performance, for our method we would like to stress that the execution of the full network will not necessarily be the bottleneck. In scenarios where the number of processors is smaller than the number of patches, scheduling will take care of evenly distributing the load. A solution to further distribute the load more evenly would be to consider multiple consecutive frames for processing. Furthermore, it is possible that no patch in an image requires full network, and the full network will never be used.

D. Framework input

The input patch size (size of patches into which the image is divided, as shown in Fig. 3) to our HVPF is equivalent to the size of the receptive field of the upsampling model employed. In the case of the VDSR network, the input patch is $\frac{40}{k} \times \frac{40}{k}$ pixels. This is due to the fact that, during the HVPF prediction, the low-resolution image is being considered. Before being conveyed to the VDSR network, the LR image is upsampled through bicubic interpolation. Consequently, in the event of $\times 4$ upsampling, the input to the HVPF is 10x10 pixels, corresponding to a patch of 40x40 pixels in the image upsampled with bicubic interpolation. In the case of the EDSR network, the input patch is 48 $\times$ 48 pixels, taken from the low-resolution image. As no prior upsampling is involved in the input image, there is no need to consider a lower-size patch. In certain instances, utilizing input patches smaller than the neural network's receptive field may be advantageous, particularly in the context of smaller images.

E. Extension - AR/VR Display

Next generation standalone virtual/augmented reality headsets demand high spatial quality, refresh-rate and power efficiency in real-time. Our framework can be applied for gaze-contingent super-resolution for AR/VR headsets. The main justification is that for wide field-of-view displays, human visual acuity decreases significantly away from the gaze-location (fovea). This inhomogeneity is frequently associated with the distribution of retinal cells across the visual field, as demonstrated by [1, 6].

Contrast sensitivity models such as the StelaCSF appropriately model human contrast perception as a function of eccentricity, and thus can extend our model to account for acuity across the visual field. Modern VR/AR headsets have built in eye-trackers that can be used to control our framework. In Fig. 3, we present some preliminary results for our quality map estimation with different gaze positions on the screen. The top row shows the eccentricity map relative to the gaze location, and the bottom row shows how our prediction for required SR quality varies. As anticipated, our perceptual model predicts that higher quality resolution will be used when the user is looking, while for areas in the periphery, our model predicts that the lowest up-sampling quality will be used. The main application is rendering in a lower resolution throughout the field of view, and then upsampling the rendering for real-time VR/AR displays using our technique.

F. More Qualitative Results

This section presents additional qualitative results, demonstrating our perceptual model predictions. Fig. 4 shows the maps generated by our method, illustrating the selective deployment of higher-quality reconstruction networks in regions of greater detail and contrast, and the use of lower-quality reconstruction networks in areas with less detail and contrast. A comparison of the SR results of the proposed method with those of the original networks is presented in Fig. 5. Furthermore, visual SR results of the proposed method and of the original network is also presented in Fig. 6,7,8,9.

References

[1] C A Curcio and K A Allen. Topography of ganglion cells in human retina. *J Comp Neurol*, 300(1):5–25, 1990. 2

[2] Jinho Jeong, Jinwoo Kim, Younghyun Jo, and Seon Joo Kim. Accelerating image super-resolution networks with pixel-level classification. In *European Conference on Computer Vision*, pages 236–251. Springer, 2024. 2

[3] Xiangtao Kong, Hengyuan Zhao, Yu Qiao, and Chao Dong. Classsr: A general framework to accelerate super-resolution networks by data characteristic. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12016–12025, 2021. 2

[4] Alexandros Stergiou and Ronald Poppe. Adapool: Exponential adaptive pooling for information-retaining downsampling, 2022. 1

[5] Shizun Wang, Jiaming Liu, Kaixin Chen, Xiaoqi Li, Ming Lu, and Yandong Guo. Adaptive patch exiting for scalable single image super-resolution. In *European Conference on Computer Vision*, pages 292–307. Springer, 2022. 2

[6] Andrew B Watson. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *J Vis*, 14(7), 2014. 2

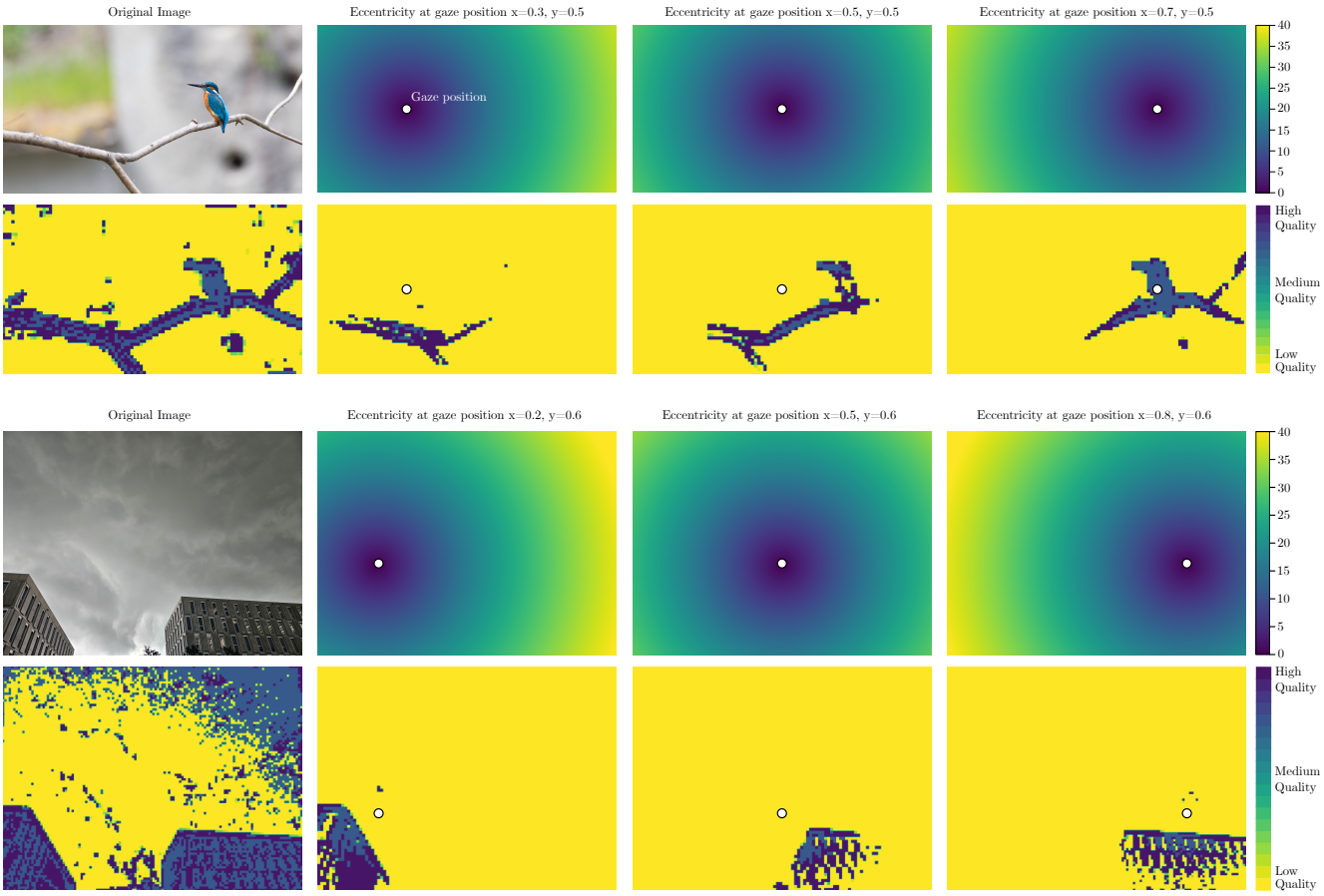

Figure 3. Our model predictions based on gaze position with $\times 4$ super-resolution. In the first column, we have the original image and the corresponding quality map. In the other columns we have on top the eccentricity map expressed in degrees and bottom we have the corresponding quality map.

[7] Wenbin Xie, Dehua Song, Chang Xu, Chunjing Xu, Hui Zhang, and Yunhe Wang. Learning frequency-aware dynamic network for efficient super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4308–4317, 2021. 2

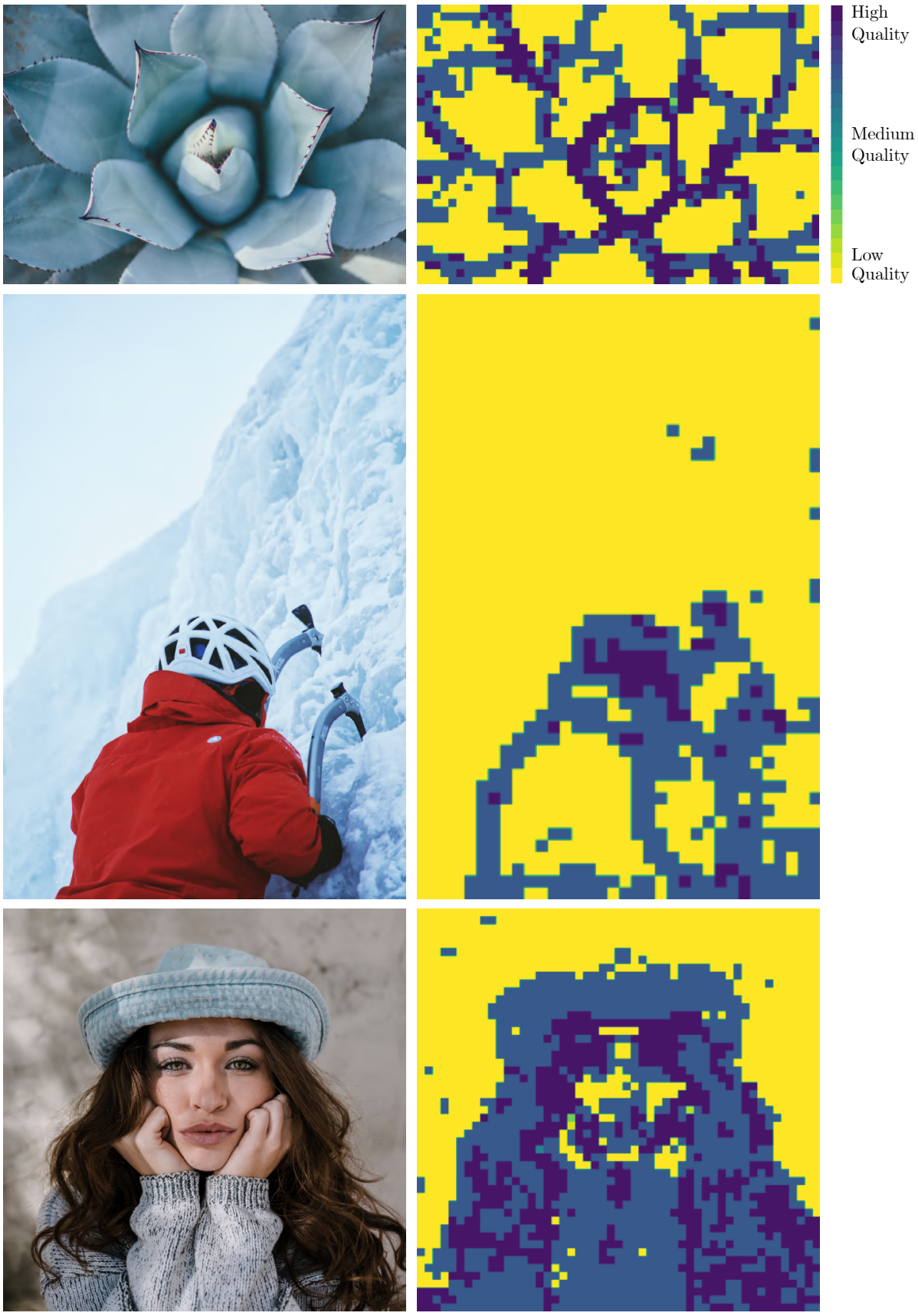

Figure 4. Our method quality map prediction results for $\times 4$ upsampling.

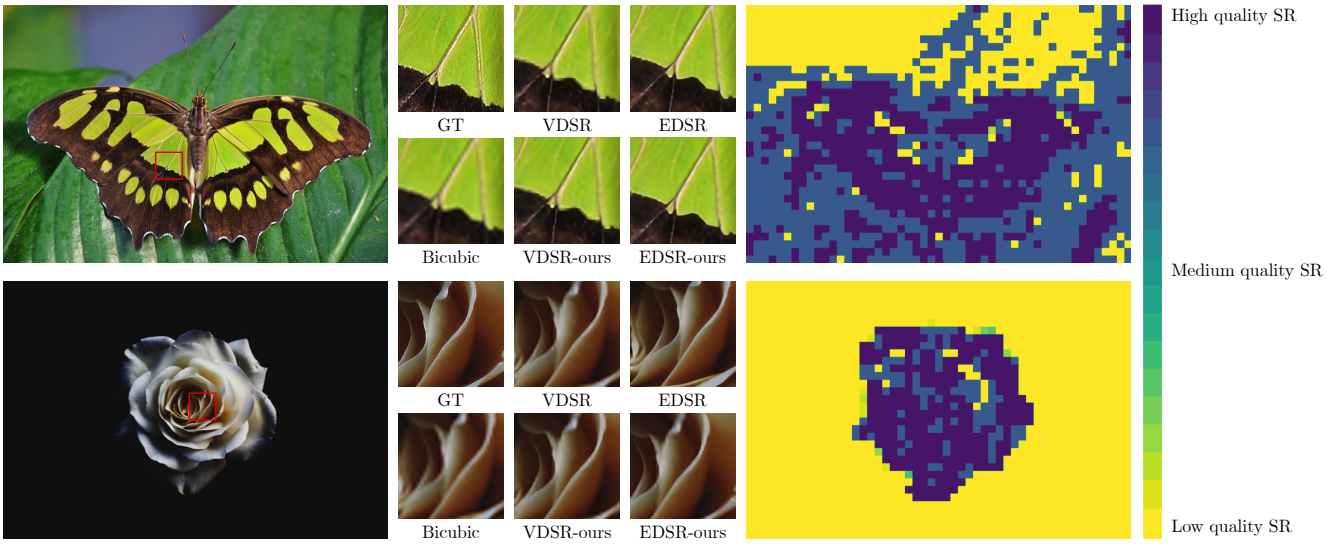

Figure 5. Visual results of our method compared to the original networks. On the right, we can observe the maps produced by our perceptual model.

(a) Original Network (b) Our method

Figure 6. The SR result ($\times 4$) of VDSR network and our method applied to VDSR network for image DIV2K-0885.

(a) Original Network

(b) Our method

Figure 7. The SR result ($\times 4$) of VDSR network and our method applied to VDSR network for image DIV2K-0878.

(a) Original Network

(b) Our method

Figure 8. The SR result ($\times 4$) of VDSR network and our method applied to VDSR network for image DIV2K-0850.

(a) Original Network (b) Our method

Figure 9. The SR result ($\times 4$) of VDSR network and our method applied to VDSR network for image DIV2K-0815.