Super Resolution for Humans

Volodymyr Karpenko
Università della Svizzera Italiana
Lugano, Switzerland
karpev@usi.ch

Taimoor Tariq
Università della Svizzera Italiana
Lugano, Switzerland
taimoor.tariq@usi.ch

Jorge Condor
Università della Svizzera Italiana
Lugano, Switzerland
jorge.condor@usi.ch

Piotr Didyk
Università della Svizzera Italiana
Lugano, Switzerland
piotr.didyk@usi.ch

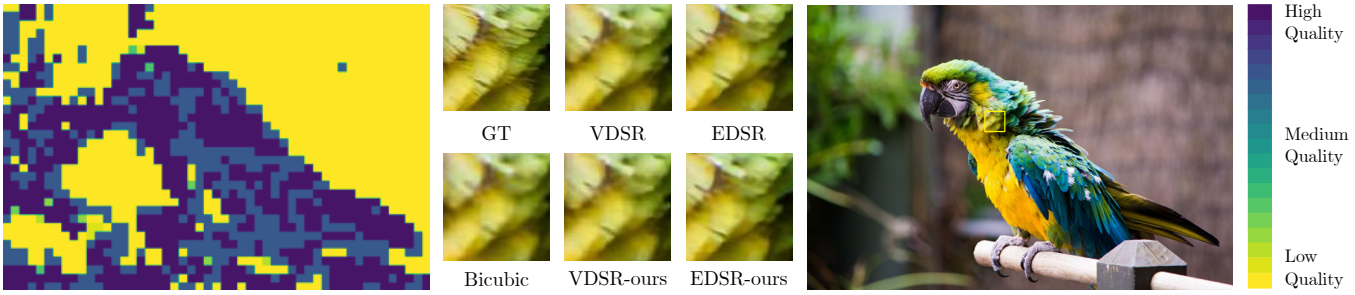

Figure 1: Our perceptually accelerated method can achieve perceptually lossless acceleration for neural network based SR.

Abstract

Super-resolution (SR) is crucial for delivering high-quality content at lower bandwidths and supporting modern display demands in VR and AR. Unfortunately, state-of-the-art neural network SR methods remain computationally expensive. Our key insight is to leverage the limitations of the human visual system (HVS) to selectively allocate computational resources, such that perceptually important image regions, identified by our low-level perceptual model, are processed by more demanding SR methods, while less critical areas use simpler methods. This approach, inspired by content-aware foveated rendering [Tursun et al. 2019], optimizes efficiency without sacrificing perceived visual quality. User studies and quantitative results demonstrate that our method achieves a reduction in computational requirements with no perceptible quality loss. The technique is architecture-agnostic and well-suited for VR/AR, where focusing effort on foveal vision offers significant computational savings.

ACM Reference Format:
Volodymyr Karpenko, Taimoor Tariq, Jorge Condor, and Piotr Didyk. 2025. Super Resolution for Humans. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH Posters '25), August 10–14, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3721250.3742985

1 Approach

Two key ideas drive our approach. First, we note that the capabilities of neural SR methods in reconstructing high spatial frequencies depend on the models' complexity and size. By creating simpler versions of the SR models, we obtained more efficient solutions but for the price of their inability to reconstruct high-frequency content. We characterize these trade-offs for a particular set of CNN models using *attenuation curves* measured on an image dataset and representing the ability of different models to reconstruct different spatial frequencies (Figure 2). The second observation is that the necessity of reconstructing high-frequency content by an SR model depends on local image content and how well a human observer can perceive the reconstructed content. For example, some image regions do not need high-quality reconstruction as their visibility is reduced due to visual masking effect. Using the two observations, we propose a perception-aware adaptive super-resolution method, which, for each image region, first analyzes the required super-resolution quality using our perceptual model and then applies the most efficient SR model that provides sufficient quality. This perceptual optimization allows us to minimize unnecessary computation that would otherwise be wasted on the reconstruction of non-perceivable spatial frequencies. In our work, we explore two ways of balancing the speed and the quality of SR solutions: network branching and different network depths. In addition, we include Bicubic interpolation as the lowest level of our hierarchy of reconstruction models for maximum efficiency. To our knowledge, this is the first attempt to optimize the super-resolution method based on the requirements of the human visual system.

Attenuation Response Estimation. To assess SR reconstruction quality, we compare the magnitude of the Fourier transform of the SR output to the ground truth for each spatial frequency. Given a ground-truth image, we downscale and then upscale it using an SR method, computing the attenuation curve:

$$\alpha_k^{\phi}(f) = \frac{1}{N}\sum_{i=1}^{N}\frac{|\mathcal{F}(\phi(I_{i\downarrow_k}))(f)|}{|\mathcal{F}(I_i)(f)|}, \tag{1}$$

where $\mathcal{F}$ is the Fourier transform, ϕ is the SR method, k the downscale factor, and N the number of images. This curve, typically in $(0, 1)$, indicates the method's ability to reconstruct each frequency

SIGGRAPH Posters '25, Vancouver, BC, Canada

ACM ISBN 979-8-4007-1549-5/25/08
https://doi.org/10.1145/3721250.3742985

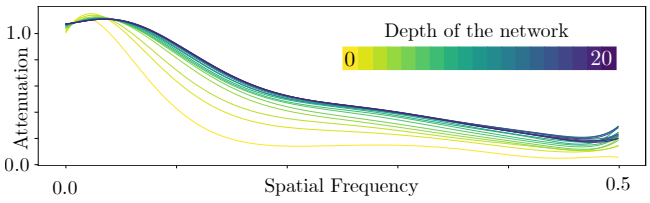

Figure 2: Attenuation curves for different network depths. Depth 0 indicates bicubic upsampling. As network depth increases, the ability to reconstruct high-frequency improves.

band. Curves are precomputed for each network variant with different k using natural image datasets. The attenuation curve serves as an efficient proxy for network performance, guiding the selection of the SR variant for each image patch.

Perceptual Contrast Modelling. Following [Tursun et al. 2019], we model local luminance contrast, $C(f,p)$, in the input image using a multi-scale Laplacian-Gaussian pyramid, where p is location and f is frequency. Contrast, $C_n(f,p)$, is normalized by the contrast sensitivity function and further adjusted for perceptual masking, yielding values in just-noticeable difference (JND) units, $C_t(f,p)$.

Perceptual optimization. For each image patch and frequency, we seek the maximal attenuation undetectable by the HVS. We define the attenuation as contrast ratio:

$$\alpha'(f) = \frac{C'(f,p)}{C(f,p)} = \frac{C'_n(f,p)}{C_n(f,p)} \tag{2}$$

where C_n,C_t and C'_n,C'_t are contrast values for input and upsampled images, respectivly. To assure that that the attenuation is undetectable, we consider additional constrain:

$$C_t(f,p) - C'_t(f,p) = 1\,JND. \tag{3}$$

By substituting the expressions for C_t and C'_t from [Tursun et al. 2019] and similarly to that work assuming that the masking term is the same in both cases, it can be shown that:

$$C'_n(f,p) = \left| |C_n(f,p)|^{0.7} - \left(1 + \sum_{q \in N(p)} \frac{|C_n(f,q)|^{0.2}}{|N|}\right)\right|^{1/0.7} \tag{4}$$

If we consider three levels of the contrast pyramid, we can compute the tolerable attenuation at selected spatial frequencies as

$$t_i = \frac{C'_n(f_i,p)}{C_n(f_i,p)}, i \in 1,2,3 \tag{5}$$

Note that t_i can be directly computed from an input image using Eq. 4 as an approximation of C'_n. Next, for each SR network branch j, we store a precomputed attenuation vector $\hat{t}_j$, SR response, computed on a image dataset. The best branch for the patch is the one whose response best matches the target attenuations for a given patch:

$$\text{branch} = \arg\max_j \left\{ \frac{t \cdot \hat{t}_j}{\|t\| \|\hat{t}_j\|} \right\} \tag{6}$$

where $t = [t_1, t_2, t_3]$ (frequencies) and j indexes the candidate branches/networks.

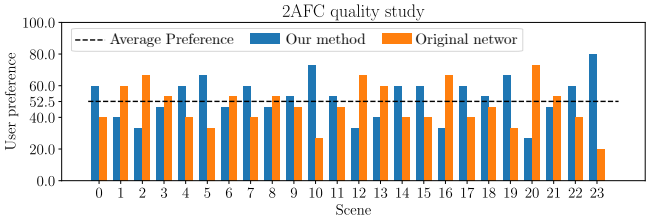

Figure 3: The result of our user study (15 subjects) for the network branching application with 24 natural images.

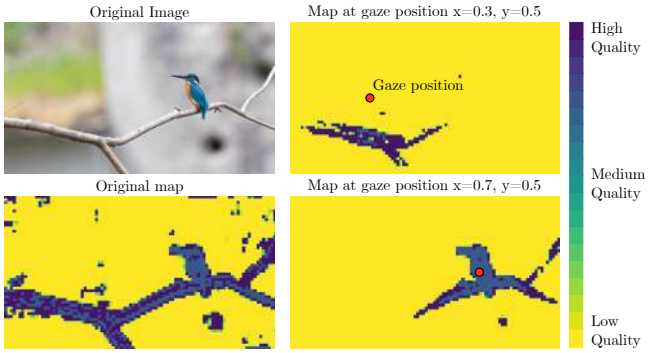

Figure 4: Our model predictions based on gaze position with ×4 super-resolution. In the first column, we have the original image and the corresponding quality map. In the other column we have the quality maps at different gaze positions.

Network branching strategies. We showcase our perceptual model to optimize SR in two settings: using early-exit branches in VDSR [Kim et al. 2016], where the optimal branch is selected per patch to balance quality and computation, and by selecting among EDSR [Lim et al. 2017] networks of different depths per patch.

2 Results and discussion

Our 2AFC user study (Fig. 3) showed that images produced by our perceptually accelerated method were indistinguishable from those of the full networks, confirming perceptually lossless acceleration. Notably, our perceptual model's computational cost is less than 1% of that required for upsampling, while still reducing FLOPs by up to 50% (VDSR) and 78% (EDSR) for ×2 upsampling, and 37% (VDSR) and 77% (EDSR) for ×4 upsampling. See Supplementary Material for further details. Integrating contrast sensitivity models like StelaCSF [Mantiuk et al. 2022], our framework enables gaze-contingent super-resolution for VR/AR, adaptively allocating high resolution only where perceptually needed (Fig. 4). Future work includes extending to video super-resolution using temporal CSF, evaluating non-CNN architectures, adapting our perceptual model to other image/video tasks (e.g., denoising, interpolation), and assessing real-world runtime beyond FLOPs.

References

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 1646–1654.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. arXiv:1707.02921 [cs.CV] https://arxiv.org/abs/1707.02921

Rafał K. Mantiuk, Maliha Ashraf, and Alexandre Chapiro. 2022. stelaCSF: a unified model of contrast sensitivity as the function of spatio-temporal frequency, eccentricity, luminance and area. 41, 4 (2022).

O. Tursun, Elena Arabadzhiyska-Koleva, Marek Wernikowski, Radosław Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk. 2019. Luminance-contrast-aware foveated rendering. *SIGGRAPH* (2019).