Human Vision Constrained Super-Resolution

Volodymyr Karpenko Taimoor Tariq Jorge Condor Piotr Didyk

{volodymyr.karpenko,taimoor.tariq,jorge.condor,piotr.didyk}@usi.ch

Università della Svizzera Italiana (USI), Lugano, Switzerland

Abstract

Modern deep-learning super-resolution (SR) techniques process images and videos independently of the underlying content and viewing conditions. However, the sensitivity of the human visual system (HVS) to image details changes depending on the underlying image characteristics, such as spatial frequency, luminance, color, contrast, or motion; as well viewing condition aspects such as ambient lighting and distance to the display. This observation suggests that computational resources spent on upsampling images/videos may be wasted whenever a viewer cannot resolve the synthesized details i.e the resolution of details exceeds the resolving capability of human vision. Motivated by this observation, we propose a human vision inspired and architecture-agnostic approach for controlling SR techniques to deliver visually optimal results while limiting computational complexity. Its core is an explicit Human Visual Processing Framework (HVPF) that dynamically and locally guides SR methods according to human sensitivity to specific image details and viewing conditions. We demonstrate the application of our framework in combination with network branching to improve the computational efficiency of SR methods. Quantitative and qualitative evaluations, including user studies, demonstrate the effectiveness of our approach in reducing FLOPS by factors of 2× and greater, without sacrificing perceived quality.

1. Introduction

Super-resolution (SR) has quickly become a fundamental tool in imaging and media distribution, given the increasing requirements of delivering higher quality content at lower bandwidths, and as general compression tools to deal with escalating imaging sensor resolutions. In media production such as virtual reality, augmented reality, and video games, SR is indispensable to cater to the high-resolution, high-framerate requirements of modern displays and low power budgets [2, 17, 33, 36]. With the advent of hardware-specific accelerators for efficiently running DL models [34], most modern approaches, even for real-time needs, involve using convolutional neural networks trained on large priors of natural images, which during run-time take low-resolution images as input and produce higher resolution versions. However, even with ever-increasing computational power at our disposal, the computational burden of high-quality SR methods is still problematic [1]. In fact, state-of-the-art methods on real-time SR using neural networks, such as Bicubic++ [8], directly compete with efficient, classic bicubic interpolation techniques, yielding marginal quantitative improvements while still struggling to compete in runtime. At the same time, the human visual system (HVS) is compressive by nature [52], meaning it has limited capabilities to resolve detail beyond some thresholds determined by viewing conditions, spatial frequencies, color or motion; any further improvement in reconstruction quality achieved by using a more expensive model can be seen as wasted resources. Our key insight is that, given that SR images are to be observed by a human, we can take advantage of these naturally compressive capabilities of the HVS [52] to process differently areas of the input image depending on its characteristics and human sensitivity to those, such as spatial frequency, luminance, color, contrast, or even, in the case of videos, motion. Computational resources are allocated to perceptually meaningful areas, as determined by our low-level visual model; analogously to how lossy compression schemes such as JPEG similarly allocate memory by leveraging the HVS.

This observation is the foundation of our work. First, we quantify the frequency reconstruction capability of baseline SR neural networks after each layer, doing so through attenuation curves. Then, after dividing an image or individual frame into a set of patches, our Human Visual Processing Framework (HVPF) (grounded on recent contrast-sensitivity functions [29]) predicts how many layers of the network should the patch be processed through; visually sensitive patches will be processed by the whole model, whereas the least visually meaningful content will simply resort to bicubic interpolation, Fig. 1 shows a visual example. We apply our model to other use cases (with the goal

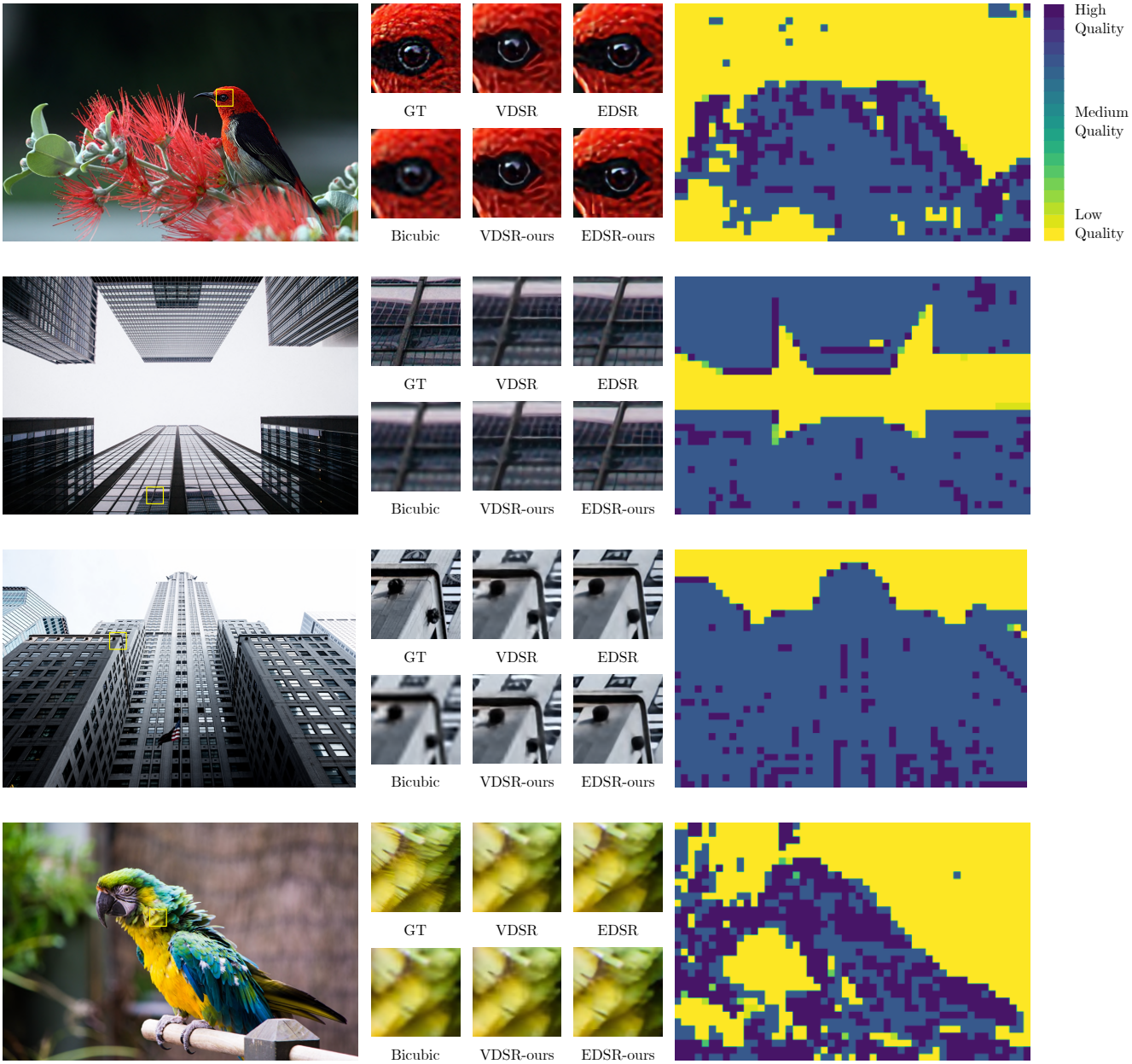

Figure 1. Visual results of our method compared to the original networks. On the right, we can observe the maps produced by our perceptual model.

of computational efficiency) such as selecting a network of appropriate depth from a set of candidates. In all of these cases, the goal is make the inference as fast as possible without noticeable quality degradation. Through a series of user studies and quantitative measurements of quality and runtime, we demonstrate that our HVPF enables faster runtime woth no perceivable loss of quality when compared with the baseline models. Furthermore, contemporary SR methods often only consider foveal vision, making them suboptimal for wide-field-of view systems such as AR/VR displays. We also propose an eccentricity-aware extension of our model for VR and AR applications, where computational efforts are further directed as per the loss of human visual acuity in peripheral vision. [36].

2. Related Work

Single-frame SR Super-Resolution or image-upsampling has traditionally been addressed through interpolation-based techniques. Nearest-Neighbor, Bilinear or Bicubic interpolation techniques are commonly used, featuring in-

creasingly bigger receptive fields and perceived quality, at the cost of performance. However, interpolation-based techniques that solely consider the input image are fundamentally limited in terms of upsampling capacity, as high frequency signals lost in the downsampled or compressed image could never be recovered from the signal itself. Leveraging natural image statistics, either explicitly [47] or implicitly [43] has shown greater potential, as lost frequencies can be composed from the expected distribution already present in natural images. Implicit (Deep-Learning based) methods have recently received the greatest attention, as large datasets of high resolution images [11, 39] and advances in neural-network architectures [13, 16, 21].

Early efforts were based on convolutional neural networks [12, 19, 45]. In practice, they learned sets of convolutional filters adapted to different features, enabling them to better reconstruct missing frequencies in a content-aware manner. These were usually trained end-to-end via downsampling HR images and upsampling the result to recover the original signal, usually employing pixel-error or perceptual [50] loss functions. A popular extension of these approaches leveraged adversarial training [14] with latent CNNs for higher quality results [44]. More recently, some methods rely on Vision Transformers [5, 13, 24, 26] for their latent architecture. The self-attention mechanism inherent in the architecture enables capturing long-range relationships within the image content, as opposed to the solely local receptive fields of CNN-based approaches. State-of-the-art methods nowadays rely on diffusion [38] rather than adversarial training [23, 46], featuring improved learning stability and quality. Despite their quality however, diffusion methods (either with vision transformer or CNN backbones) are computationally expensive and usually disregarded in applications where performance is primed.

Temporally-consistent Video SR Temporally-consistent video SR has received ample attention in the literature, for both pre-rendered video [22] and real-time content (i.e. videogames) [2, 17, 33]. The main difference between single image SR is the availability of additional frames, as well as extra information from the rendering engine in the case of videogames (i.e. motion vectors, material information). Traditionally, single-image learning-based approaches struggle to provide consistent SR across frames due to the implicit, difficult-to-interpret latent space they leverage, which does not guarantee temporally consistent outputs. To ensure smoothness across frames, most works simply condition the current frame upsampling on previous or subsequent frames through motion vectors or optical flow to ensure smoothness [33].

There is a fundamental difference between the traditional approach to video SR and our proposal that is important to clarify. Traditionally, video SR methods aim to achieve a better spatial reconstruction due to the availability of multiple frames. However, human sensitivity to spatial details decreases with movement, so a human vision centric approach should leverage this and reduce quality as a function of movement magnitude, without noticeable visual degradation. Our proposed model quantifies the loss of visual acuity with motion, and appropriately decreases the spatial quality of SR. Therefore, we can get faster per-frame SR with videos, in a manner adaptive to factors such as the nature of the content and amount of movement.

3. Background

3.1. Low-Level Human Vision

Owing to evolutionary imperatives of efficiency, the HVS has evolved to be a compressive system [52]. What this essentially means is that we are not able to resolve all the visual information that is coming into our eyes. The first critical aspect is that due to center surround receptive field of the early visual system, we are much more sensitive to variations in contrast rather than absolute luminance [40]. Furthermore, it is well known that human contrast perception is highly dependent on factors such as spatial frequency and luminance. Due to the nature of the collective receptive field our early human vision, we are most sensitive to a narrow band of spatial frequencies, and our sensitivity falls of at lower and higher spatial frequencies. This phenomena is aptly captured by a model of the early visual system called the Contrast Sensitivity Function (CSF) [6]. The CSF is highly dependent on factors such as local adapting luminance, size of the stimulus, eccentricity, and the amount of movement. For example, universal sensory models such as the Weber-Fechner law tell us that our ability to detect contrast decreases with increasing luminance [27]. Furthermore, due to effects such as contrast masking, neighboring contrast is known to strongly effect human visual perception [41]. This is why we are sometimes less likely to see a loss of resolution in highly textured areas as opposed to independent strong edges. Another very important aspect is that movement or temporal variation decreases our ability to detect and resolve contrast [4]. This behavior is quantified by the dependency of the CSF on temporal frequencies as well as spatial frequencies.

3.2. Visual Difference Predictors (VDP)

Inspired by the compressive nature of the human vision, Visual Difference Predictors (VDP) are models that aim to predict the perceived differences between two images based on robustly modeling the frequency selective nature of the early visual system. The framework was originally introduced by Daly et al. [10]. Since then, there have a lot of improvements and extensions to the original model. [27] designed the HDR-VDP, which was one of the first human

perception inspired metrics aim to quantify visual differences between HDR images. [28] then extended the VDP to account for human peripheral vision and color [30]. [42] employed the VDP framework to control spatial quality in VR-HMDs. [41] designed a variation of the VDP framework for real-time perceptually optimized tone mapping. Our HVPF can be thought of as a VDP framework specifically tailored for real-time application to the problem of Deep Learning based SR. To the best of our knowledge, our framework is the first application of robust Human Vision frameworks to efficient neural network based image/video processing.

4. Our Method: A Human Visual Processing Framework (HVPF)

Our approach is centered around the frequency domain interpretation of SR and the well-established fact that the early visual system is frequency-selective. The main motivation is that the difference between low and high-resolution images lies in the attenuation and removal of higher spatial frequencies. The task of an SR neural network is to reconstruct the missing high spatial frequencies. The better and stronger the neural network, the better the reconstruction. The main idea behind our work is that due to the limitations of the HVS, we do not always need a perfect reconstruction. If we can quantify the spectral nature of an SR method, we can guide the method using models of human visual perception to deliver the least expensive reconstruction required for optimal visual quality, i.e., any further improvement in reconstruction leads to no perceivable benefit while wasting computational resources.

While there are various ways to control the trade-off between the computational efficiency of an SR method and the reconstruction quality, we consider two: network branching and altering network depth. The first approach adds earlier exit points to the original network. Using earlier exit points, i.e., shallower branches, leads to less computation and lower reconstruction quality. In the latter method, different variants of an SR method are created by varying the depth of the original network to make shallower networks more efficient yet potentially comprising the visual quality of the output. Our method is not limited to the above techniques, and others, such as network quantization, could be easily incorporated.

Given different variants of an SR method, our method aims to predict which version should be used in a specific region of an image or video frame. We propose to use attenuation curves to first quantify the reconstruction capability of a given variant. The attenuation curves express the ratio of the radially averaged 2D Fourier transform of the reconstructed output and its full-resolution counterpart. We demonstrated that such curves can be computed on a set of images and reused. Furthermore, we design a framework that expresses the required reconstruction quality in the form of the attenuation curve. Later, our method selects an appropriate SR variant to ensure adequate reconstruction quality for a region while minimizing computational costs. In practice, our method works on image patches which are both input to the SR method and our prediction. Below, we describe all the components of the method.

4.1. Attenuation Response Estimation

For a given SR method and an input image, we can quantify the quality of reconstruction by comparing the magnitude of the Fourier Transform of the reconstructed image and its ground-truth version. Given a ground-truth image I, we first downscale it by a factor k, producing image $I{\downarrow_k}$. Later, we use a SR method to upscale that image back to its original resolution. Comparing the Fourier transform of the resulting image $\hat{I}$ and the ground-truth counterpart I, allows us to quantify the reconstruction power of the analyzed SR method. More formally, given a SR method ϕ and a test image I, we define the frequency dependent attenuation curve as:

$$\alpha_k^{\phi}(I, f) = \frac{|\mathscr{F}\left(\phi(I{\downarrow_k})\right)(f)|}{|\mathscr{F}\left(I\right)(f)|}, \tag{1}$$

where $\mathscr{F}$ denotes the Fourier transform. Since we are interested in characterizing the reconstruction capability of the method, we do not compute the curve for one image but for a set of images $\{I_k\}$ and compute for a given SR method ϕ and downscaling factor k the aggregated attenuation response curve as an average across all the images:

$$\alpha_k^{\phi}(f) = \sum_{i=1}^{N} \alpha_k^{\phi}(I_i, f). \tag{2}$$

Although not guaranteed, the value of α_k^{ϕ} is expected to lie on $(0, 1)$ range, where $\alpha_k^{\phi}(f) = 0$ means that the SR method was not able to reconstruct content at spatial frequency f, while $\alpha_k^{\phi}(f) = 1$ indicates the full capability in reconstructing this part of the signal. It has to be noted that this measure quantifies the presence of the signal in the reconstructed output and not its correctness. Nevertheless, we use this as a proxy for the reconstruction quality. This choice was further motivated by the fact that, although neural networks are non-linear functions, SR is a low-level task with consistent and deterministic characteristics in the frequency domain. The average spectrum of natural images is known to adhere to consistent Fourier characteristics like the inverse power law fall-off. Thus, an average attenuation curve is a good approximation for network response to a general natural image for the task of SR. In order to be more conservative, there is a possibility to use a particular percentile or quartile around the average, but our empirical analysis did not reveal it necessary, which is later confirmed by the results of our user experiments.

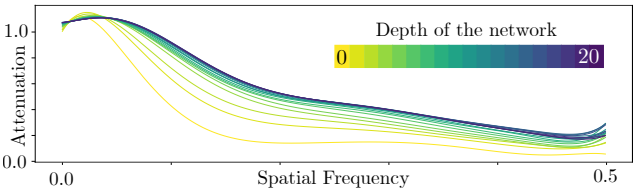

Figure 2. The attenuation curves derived for the case of modulating the depth (number of layers) of the network. Depth equal 0 corresponds to a bicubic upsampling. As the depth of the network is reduced the performance of the solution in reconstructing high-spatial-frequency signal is reduced.

For all the variants of the SR techniques considered in this work, we pre-compute the attenuation response curves using the above procedure. We always use 19 natural images from the set5 [7] and set14 [49] datasets, and compute different sets of curves for downscaling parameter $k \in \{2, 4, 8\}$. Fig. 2 presents a set of curves for the case of varying the network depth. To obtain a more compact representations of the attenution functions, we model them using Gaussian fall-off:

$$\alpha'(f) = \frac{1}{a\sqrt{2\pi}} \exp\left(-\frac{(f-b)^2}{2a^2}\right) + c, \tag{3}$$

where f is the spatial frequency and a, b, c are parameters estimated via fitting. After estimating the attenuation curves for variants of SR methods, the next step is to estimate which of them is ideal for a given patch of the image or video frame.

4.2. Perceived Contrast Modeling

We model the perceived luminance contrast following the VDP framework e.g. [42]. Please refer to the original work for more details on perceptual difference modeling. To summarize our model designed for application to SR, luminance contrast is computed as $C(f, p)$ using a multi-scale Laplacian-Gaussian pyramid, where p is the location and f is spatial frequency [9, 37]. The contrast measure is then normalized by the contrast sensitivity function (CSF), yielding $C_n(f, p)$. Finally, the perceived contrast, $C_t(f, p)$, also incorporates the visual masking model [48] with parameters $\alpha = 0.7$ and $\beta = 0.2$.

4.3. Optimization

Given an input patch, the goal is to find the maximum attenuation that is under the resolution capabilities of human vision. According to the contrast model, the attenuation remains undetectable by an observer if the contrast difference between the original image and the attenuated one is under 1 JND. Consequently, to find the attenuation that results in maximum performance gains yet imperceptible quality loss, we optimize the attenuation curve such that it results in exactly 1 JND difference, i.e.,

$$\forall_f \; C'_t(p, f) - C_t(p, f) = 1, \tag{4}$$

where $C_t(p, f)$ represents the perceived contrast of the input image patch and $C'_t(p, f)$ represents the perceived contrast of the network output. Assuming the attenuation curves are a response of the network/branch to the input, the attenuation is a modulation of the physical image contrast at different frequencies:

$$\alpha'(f) = \frac{C'(f, p)}{C(f, p)} = \frac{C'_n(f, p)}{C_n(f, p).} \tag{5}$$

Our immediate goal is to estimate the tolerable output contrast $C'_n(f, p)$ from the constraint in Eq. (4). We start by developing Eq. (4) by substituting expressions from [42], which leads to the following form of the constraint:

$$\frac{\text{sign}\left(C'_n(f,p)\right) \cdot \left|C'_n(f,p)\right|^{\alpha}}{1 + \frac{1}{|N|} \sum_{q \in N(p)} |C'_n(f,q)|^{\beta}} - \frac{\text{sign}(C_n(f,p)) \cdot |C_n(f,p)|^{\alpha}}{1 + \frac{1}{|N|} \sum_{q \in N(p)} |C_n(f,q)|^{\beta}} = 1, \tag{6}$$

where the numerators encode the CSF-weighted contrast values, while the denominators model visual masking effect. It is evident that $C'_n(f, q)$ cannot be directly calculated from this equation due to the visual masking term present in the denominator. To address this problem, similarly to the work of [42], we make the assumption that the contrast masking for the up-sampled output patch can be approximated by that of the input patch. Furthermore, knowing that the sign of the contrast should remain unaltered throughout the neural network processing Eq. (6) we can derive $C'_n(f, p)$ directly as:

$$C'_n(f, p) = \left| \left(1 + \textstyle\sum_{q \in N(p)} \frac{|C_n(f,q)|^{\beta}}{|N|}\right) - |C_n(f, p)|^{\alpha} \right|^{1/\alpha} \tag{7}$$

It should be noted that the sign of the contrast is omitted since the focus is on the magnitude of the contrast itself. Now assuming that we have three levels of the contrast pyramid, tolerable attenuation at three different spatial frequencies can be calculated as follows:

$$t_i = \frac{C'_n(f_i, p)}{C_n(f_i, p)}, i \in 1, 2, 3 \tag{8}$$

It is important to note that thanks to the derivation in Eq. (7), the attenuation can be computed directly from C_n, i.e., the input patch only.

Having the tolerable attenuation $t = \{t_1, t_2, t_3\}$, we can find the most suitable SR network/branch by identifying the one with most similar attenuation curve (Section 4.1, Eq. (3)). More formally, this step can be defined using following optimization problem:

$$\text{branch/network} = \arg\max_j \left\{ \frac{t \cdot \hat{t}_j}{\|t\| \|\hat{t}_j\|} \right\}, \tag{9}$$

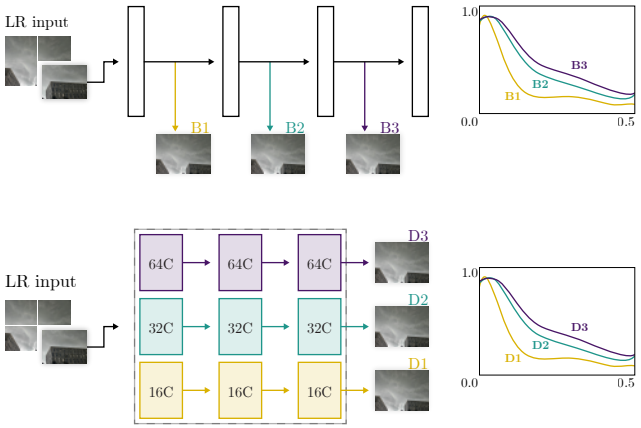

Figure 3. Flowchart illustrating the methodology employed for efficient SR. The input image is divided into patches. For each patch, our task is to estimate which branch B (for the branching case) and which network D (for the depth case) should be used to each patch. The estimation is made through analyzing the computed attenuation characterstics (right). Such as to minimize computations with noticeable quality degradation.

where $\hat{t}_j = \{\alpha'(f_1), \alpha'(f_2), \alpha'(f_3)\}$ is the vector of the estimated attenuation produced by Eq. (3) for a given netwrok/brach j.

4.4. Model Efficiency

In our application, it is essential that the overhead of the HVPF is minimal compared to the performance improvements provided by the SR method. Although we prototyped our solution using Python and evaluated it on a single CPU thread, it has been shown that a similar processing pipeline can be implemented efficiently on a GPU. This is primarily due to the nature of the computation, which consists of independent per-pixel operations (see Eq. (7)). Additionally, the construction of the contrast pyramid can be accelerated using MIPMAP functionality. Rencently, [41] have demonstrated a HVPF on the Meta Quest 2 VR headset, achieving 2K resolution with a single execution time of under 1 ms.

5. Experimental Setup

Validation on SR Models As mentioned in Section. 4, we employ our HVPF to optimize SR for two cases. The first approach adds earlier exit points (branches) to the original network. Using earlier points, i.e., shallower branches, leads to less computation and lower reconstruction quality. To test this case, we employ the popular and seminal VDSR [20] network. A neural network with 19 branched outputs was created as per the setting in Fig. 3 (top). After each ReLU activation function, an exit point was added, structured identically to the final layer of the original network., similar to [18]. Our task was to to use the HVPF to select the appropriate branch (per image patch) such that there is no noticeable quality loss.

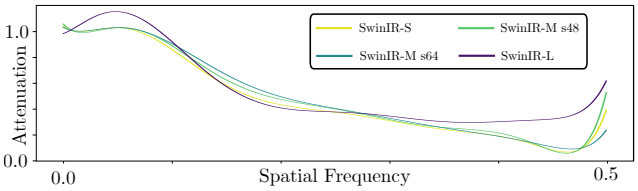

Figure 4. Attenuation curves derived from the SwinIR, a transformer based SR model. As the size of the model increases the capability of reconstructing higher frequency details increases.

The second approach is reducing the depth or number of channels of a network to make it more efficient. In this case, we have a number of candidate networks with different number of channels-per-layer, and our task is to employ our HVPF to select the appropriate one per image patch. To test this case, we employ the EDSR [25] network. In the case of EDSR, five networks with varying numbers of channels per layer (256, 128, 64, 16, 8) were trained independently, the training procedure was the same as described in [25]. Our task was to use the HVPF to select the appropriate network (per image patch) such that there is no noticeable quality loss. We include additional details on our choice of patch sizes in Sec. D.

On the generality of our visual framework. We selected EDSR and VDSR models as test benchmarks for HVPF due to their widespread use, robustness and demonstrated efficacy over the years. Our algorithm is network agnostic: most SR approaches will follow similar attenuation characteristics, which are grounded in the nature of natural images and the Fourier nature of the SR problem [30]. In Fig. 4 we show attenuation curves for Transformer-based models, where increasing model size directly correlates with its ability to reconstruct higher spatial frequency content. This is also in line with previous research on implicit models for vision and graphics, where increasing the number of weights directly correlates with the capacity of the model to learn higher frequency content [31] and together with our results on CNNs makes us confident on the generality of our HVPF to be leveraged with any SR method.

Subjective Quality Study. We developed a perceptual experiment with human subjects in order to validate our approach.

Task. We employed a 2AFC experimental protocol. The task was a forced choice between two test cases in relation to a given high quality reference. Users were instructed to select the test case which was perceived more similar to the reference. On the right side we display the high quality reference; while the left side displayed either

- A) our HVPF-powered SR, selecting the appropriate network/branch for the corresponding image patch or,
- B) The output of the unaltered full deep network applied to the whole image.

These were displayed in randomized order. Users could use the space-bar to switch between A and B, and press ENTER when they had made their choice.

Stimuli. We employed 24 high-quality natural scenes for our user study. The scenes were selected for diversity in characteristics such as luminance, contrast, and texture. Each scene was downscaled by a factor of $\times 4$ and upscaled back to the original resolution.

In Sec. A we included additional details on the experimental setup.

6. Results and Discussion

To evaluate the effectiveness of our method, we tested the model on images and videos, as our model is additionally capable of handling the temporal frequencies present in videos.

6.1. Quantitative Results

The quantitative results for the image datasets are presented in Tab. 1. The proposed method allows for comparable performance to the original networks while reducing the computational cost. For instance, the $\times 2$ and $\times 4$ upsampling operations exhibit a reduced computational cost ranging from 58% to 22% and from 70% to 20%, respectively. For a single patch of size 10×10 the computational cost of our method is 39KFlops, while for a patch size of 35×35 is 477KFlops.

Greater savings were achieved with $\times 2$ upsampling in comparison to $\times 4$ upsampling, which is in line with expectations. This is due to the increased presence of high-frequency information in the images, which allows for reconstructing certain parts of the image with less computational power.

In order to evaluate the video, we estimated the optical flow. Subsequently, we calculated the temporal frequency that was necessary for our model, based on the velocities obtained with the optical flow. The frame rate considered for each video was 24 fps. The results for the video datasets are presented in Tab. 2.

In Sec. C we provide an explanation for the choice of FLOPS as a measure of efficiency.

6.2. Subjective Quality Results

It is well-known that metrics such as PSNR, SSIM, etc are not correlated with human quality assessments, unable to model the intricacies of how humans perceive image quality [15, 32, 35]. As our HVPF is based on a robust and detailed modeling of early visual processing in a scene dependent manner (unlike heuristic metrics trained over a large set of images e.g [51]), our hypothesis (little noticeable quality loss) can only be aptly verified using subjective quality studies.

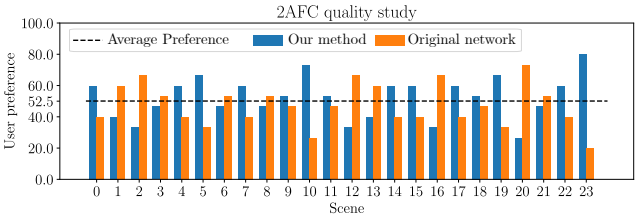

Figure 5. The result of our subjective study (for 15 participants) for the network branching application.

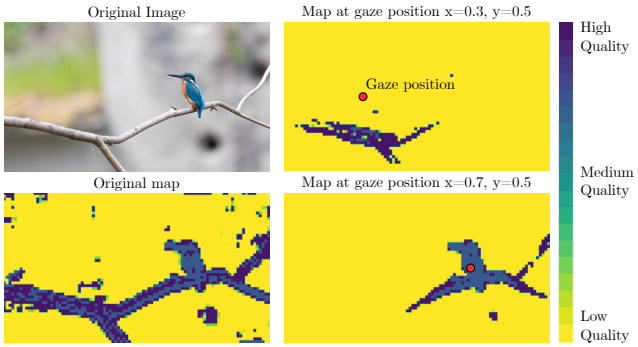

Figure 6. Our model predictions based on gaze position with $\times 4$ super-resolution. In the first column, we have the original image and the corresponding quality map. In the other column we have the quality maps at different gaze positions.

Fig. 5 shows the results of our 2AFC user study on images for the network branching application case on the VDSR. It can be seen that on average, the preference value hovers around 50%, which is indicative that users were not able to perceive any difference between the test cases A and B in relation to the reference, which supports our initial goal and hypothesis. As expected, no differences were perceived even with the reported differences in PSNR and SSIM in Tab. 1. Further results for the study on videos are provided in Sec. B.2, which demonstrates the efficacy of our HVPF for handling motion too.

A more detailed study on the network channel depth SR alternative (EDSR, bottom Fig. 3) is presented in Sec. B.1, providing similar conclusions. In summary, the studies demonstrate that the results indeed conform with our hypothesis that there is no perceivable loss in visual quality, even when there are significant computational savings through the application of our HVPF.

Finally, we also include promising early results on foveated SR (super resolution aware of gaze position, leveraging the substantially degraded visual acuity on the peripheral vision) in Fig. 6 and in Sec. E.

Table 1. **Quantitative comparison on image datasets**

Method	Scale	Set5			Set14			BSD100			Urban100			DIV2K		
		PSNR↑	SSIM↑	FLOPS↓	PSNR↑	SSIM↑	FLOPS↓	PSNR↑	SSIM↑	FLOPS↓	PSNR↑	SSIM↑	FLOPS↓	PSNR↑	SSIM↑	FLOPS↓
Bicubic	×2	32.32	0.923		28.60	0.859		28.22	0.834		25.48	0.840		31.23	0.898	
VDSR	×2	34.15	0.946	89.60G (100%)	29.98	0.899	173.68G (100%)	29.68	0.885	131.04G (100%)	27.17	0.900	144.99G (100%)	32.65	0.931	2.04T (100%)
VDSR-ours	×2	32.53	0.936	**51.99G (58%)**	29.29	0.895	**100.18G (57%)**	29.40	0.883	**61.63G (47%)**	26.48	0.889	**85.43G (58%)**	32.08	0.928	**1.06T (51%)**
EDSR	×2	36.42	0.954	1.54T (100%)	32.03	0.905	2.97T (100%)	30.62	0.888	2.25T (100%)	30.42	0.932	2.56T (100%)	34.84	0.939	31.64T (100%)
EDSR-ours	×2	36.00	0.951	**353.09G (22%)**	31.58	0.901	**691.09G (23%)**	30.26	0.882	**466.39G (20%)**	29.74	0.925	**606.62G (23%)**	34.25	0.925	**7.27T (22%)**
Bicubic	×4	26.98	0.790		24.28	0.676		24.54	0.638		21.89	0.642		26.80	0.756	
VDSR	×4	28.13	0.827	89.60G (100%)	25.01	0.709	173.68G (100%)	25.11	0.670	131.04G (100%)	22.75	0.698	572.38G (100%)	27.52	0.786	2.04T (100%)
VDSR-ours	×4	27.77	0.823	**64.13G (71%)**	24.86	0.713	**118.64G (68%)**	25.08	0.670	**69.27G (52%)**	22.65	0.699	**387.01G (67%)**	27.44	0.788	**1.29T (63%)**
EDSR	×4	30.60	0.878	579.49G (100%)	26.95	0.753	1.00T (100%)	26.01	0.706	695.39G (100%)	24.82	0.776	3.16T (100%)	29.05	0.822	10.32T (100%)
EDSR-ours	×4	30.27	0.874	**145.34G (25%)**	26.69	0.750	**236.71G (23%)**	25.82	0.701	**173.54G (24%)**	24.48	0.765	**746.13G (23%)**	28.75	0.816	**2.46T (23%)**

Table 2. **Quantitative comparison on video datasets X4 upscaling**

Method	REDS			Vid4			UDM10		
	PSNR↑	SSIM↑	FLOPS↓	PSNR↑	SSIM↑	FLOPS↓	PSNR↑	SSIM↑	FLOPS↓
Bicubic	26.39	0.724		22.44	0.614		30.76	0.884	
VDSR	27.11	0.756	702.24G (100%)	23.14	0.670	291.96G (100%)	31.71	0.899	680.96G (100%)
VDSR-ours w/o temporal frequency	27.03	0.755	443.84G (63%)	23.06	0.667	194.23G (66%)	31.62	0.903	433.30G (63%)
VDSR-ours	26.66	0.739	**173.77G (24%)**	23.05	0.667	**193.00G (66%)**	31.58	0.902	**386.77G (56%)**
EDSR	28.27	0.791	3.24T (100%)	23.92	0.711	1.59T (100%)	34.28	0.929	3.24T (100%)
EDSR-ours w/o temporal frequency	27.03	0.755	813.12G (25%)	23.73	0.703	**348.84G (21%)**	33.78	0.924	813.96G (25%)
EDSR-ours	27.03	0.755	**586.58G (18%)**	23.73	0.703	**348.84G (21%)**	33.84	0.925	**801.61G (24%)**

7. Limitations and Future work

Although the method demonstrated satisfactory performance with regard to video content, it was observed that aliasing issues were present. This phenomenon is common when individually processing frames as isolated still images. In addition, our method does not explicitly guarantee the spatial consistency between the patches. However, no such inconsistencies were noted or reported by the participants in our experiments. It would be of interest to investigate the potential of our HVPF in conjunction with video-specific SR techniques. Evaluating the performance of our model within foveation-aware frameworks is also an interesting avenue for future work, given that our choice of contrast-sensitivity function (StelaCSF) is capable of modelling contrast sensitivity based on eccentricity. Finally, our estimation is based solely on image luminance; we could further leverage additional information, such as colour, by integrating recent advances in color-aware CSFs like CastleCSF [3].

8. Conclusions

We present a thorough novel framework to leverage the specific deficiencies of the HVS to optimize computational resources in the context of SR. HVPF is fast, robust, and can be seamlessly integrated into any SR framework, adaptively optimizing computational resources to the areas in the image that require it, from the assumption that a human will be the final observer of the resulting SR image. While normally SR methods are evaluated in terms of reconstruction quality, through metrics such as PSNR or SSIM, these metrics do not model the intricacies of human visual quality assessment, and have been repeatedly demonstrated as uncorrelated with human visual perception [15, 32, 35]. In contrast, we validate our framework through a series of human studies, showcasing indistinguishable quality at substantially reduced computational costs. Our results demonstrate that even with FLOP reductions by factors of-and-greater than 3x, our HVPF minimally degrades the SR output such that the final result is not visually distinguishable from the output of a fully-resourced deep network. We are confident that our method can be further extended in the future to better integrate with video-specific SR approaches, and will be particularly relevant on VR and AR eccentricity-aware SR frameworks, where computational savings can be pushed significantly further.

Acknowledgement

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grand agreement No 804226).

References

[1] Adobe Inc. Lightroom super resolution. https://www.adobe.com/products/photoshop-lightroom/

super-resolution.html, 2024. Accessed: 2024-11-15. 1

[2] AMD Corporation. Fidelityfx super resolution (fsr). https://www.amd.com/en/technologies/radeon-software-fidelityfx-super-resolution, 2024. Accessed: 2024-11-15. 1, 3

[3] Maliha Ashraf, Rafał K. Mantiuk, Alexandre Chapiro, and Sophie Wuerger. castlecsf — a contrast sensitivity function of color, area, spatiotemporal frequency, luminance and eccentricity. *Journal of Vision*, 24(4):5–5, 2024. 8

[4] Maliha Ashraf, Rafał K. Mantiuk, Alexandre Chapiro, and Sophie Wuerger. castlecsf — a contrast sensitivity function of color, area, spatiotemporal frequency, luminance and eccentricity. *Journal of Vision*, 24, 2024. 3

[5] Neeraj Baghel, Shiv Ram Dubey, and Satish Kumar Singh. Srtransgan: Image super-resolution using transformer based generative adversarial network. *ArXiv*, abs/2312.01999, 2023. 3

[6] Peter G. J. Barten. Formula for the contrast sensitivity of the human eye. In *IS&T/SPIE Electronic Imaging*, 2003. 3

[7] Marco Bevilacqua, Aline Roumy, Christine M. Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British Machine Vision Conference*, 2012. 5

[8] Bahri Batuhan Bilecen and Mustafa Ayazoglu. Bicubic++: Slim, slimmer, slimmest-designing an industry-grade super-resolution network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1623–1632, 2023. 1

[9] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4): 532–540, 1983. 5

[10] Scott J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Electronic imaging*, 1992. 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:295–307, 2014. 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 3

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020. 3

[15] Param Hanji, Rafal Mantiuk, Gabriel Eilertsen, Saghi Hajisharif, and Jonas Unger. Comparison of single image hdr reconstruction methods — the caveats of quality assessment. In *ACM SIGGRAPH 2022 Conference Proceedings*, New York, NY, USA, 2022. Association for Computing Machinery. 7, 8

[16] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 3

[17] Intel Corporation. Xe super sampling (xess). https://www.intel.com/content/www/us/en/products/docs/discrete-gpus/arc/technology/xess.html, 2024. Accessed: 2024-11-15. 1, 3

[18] Dohyun Kim, Joongheon Kim, Junseok Kwon, and Tae-Hyung Kim. Depth-controllable very deep super-resolution network. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. 6

[19] Jiwon Kim, Jung Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. 2015. 3

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016. 6

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc. 3

[22] Gen Li, Jie Ji, Minghai Qin, Wei Niu, Bin Ren, Fatemeh Afghah, Lin Guo, and Xiaolong Ma. Towards high-quality and efficient video super-resolution via spatial-temporal data overfitting. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10259–10269, 2023. 3

[23] Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhi hai Xu, Qi Li, and Yue ting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2021. 3

[24] Jingyun Liang, Jie Cao, Guolei Sun, K. Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021. 3

[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution, 2017. 6

[26] Zhisheng Lu, Juncheng Li, Hong Liu, Chao Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 456–465, 2021. 3

[27] Rafał K. Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. *SIGGRAPH*, 2011. 3

[28] Rafał K. Mantiuk, Alexandre Chapiro, Gizem Rufo, Trisha Lian, Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro,

and Anton Kaplanyan. Fovvideovdp: A visible difference predictor for wide field-of-view video. *SIGGRAPH*, 40:1 – 19, 2021. 4

[29] Rafał K. Mantiuk, Maliha Ashraf, and Alexandre Chapiro. stelacsf: a unified model of contrast sensitivity as the function of spatio-temporal frequency, eccentricity, luminance and area. *ACM Trans. Graph.*, 41(4), 2022. 1

[30] Rafał K. Mantiuk, Param Hanji, Maliha Ashraf, Yuta Asano, and Alexandre Chapiro. Colorvideovdp: A visual difference predictor for image, video and display distortions. *SIGGRAPH*, 2024. 4, 6

[31] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2021. 6

[32] Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020. 7, 8

[33] NVIDIA Corporation. Deep learning super sampling (dlss). https://www.nvidia.com/en-us/geforce/technologies/dlss/, 2024. Accessed: 2024-11-15. 1, 3

[34] NVIDIA Corporation. Nvidia tensor core architecture. https://www.nvidia.com/en-us/data-center/tensorcore/, 2024. Accessed: 2024-11-15. 1

[35] Jean-François Pambrun and Rita Noumeir. Limitations of the ssim quality metric in the context of diagnostic imaging. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2960–2963, 2015. 7, 8

[36] 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 Trans. Graph.*, 35(6), 2016. 1, 2

[37] E Peli. Contrast in complex images. *J Opt Soc Am A*, 7(10): 2032–2040, 1990. 5

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 3

[40] Robert Shapley, Ehud Kaplan, and Keith P. Purpura. Contrast sensitivity and light adaptation in photoreceptors or in the retinal network. *Contrast Sensitvity (Proceedings of the Retina Research Foundation Symposia)*, 1993. 3

[41] Taimoor Tariq, Nathan Matsuda, Eric Penner, Jerry Jia, Douglas Lanman, Ajit Ninan, and Alexandre Chapiro. Perceptually adaptive real-time tone mapping. *SIGGRAPH Asia*, 2023. 3, 4, 6

[42] O. Tursun, Elena Arabadzhiyska-Koleva, Marek Wernikowski, Radosław Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk. Luminance-contrast-aware foveated rendering. *SIGGRAPH*, 2019. 4, 5

[43] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10576–10585, 2021. 3

[44] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018. 3

[45] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. A fully progressive approach to single-image super-resolution. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 977–97709, 2018. 3

[46] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xing Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13049–13059, 2023. 3

[47] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 3

[48] Wenjun Kevin Zeng, Scott J. Daly, and Shawmin Lei. Point-wise extended visual masking for jpeg-2000 image compression. *International Conference on Image Processing (ICIP)*, 1:657–660 vol.1, 2000. 5

[49] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, pages 711–730, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 5

[50] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 3

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[52] Li Zhaoping. Theoretical understanding of the early visual processes by data compression and data selection. *Network: computation in neural systems*, 17(4):301–334, 2006. 1, 3