



Learning GAN-based Foveated Reconstruction to Recover Perceptually Important Image Features

LUCA SURACE, Università della Svizzera italiana, Switzerland

MAREK WERNIKOWSKI, Università della Svizzera italiana, Switzerland and West Pomeranian University of Technology, Poland

CARA TURSUN, Università della Svizzera italiana, Switzerland and University of Groningen, Netherlands

KAROL MYSZKOWSKI, Max Planck Institute for Informatics, Germany

RADOSŁAW MANTIUK, West Pomeranian University of Technology, Poland

PIOTR DIDYK, Università della Svizzera italiana, Switzerland

A foveated image can be entirely reconstructed from a sparse set of samples distributed according to the retinal sensitivity of the human visual system, which rapidly decreases with increasing eccentricity. The use of Generative Adversarial Networks has recently been shown to be a promising solution for such a task, as they can successfully hallucinate missing image information. As in the case of other supervised learning approaches, the definition of the loss function and the training strategy heavily influence the quality of the output. In this work, we consider the problem of efficiently guiding the training of foveated reconstruction techniques such that they are more aware of the capabilities and limitations of the human visual system, and thus can reconstruct visually important image features. Our primary goal is to make the training procedure less sensitive to distortions that humans cannot detect and focus on penalizing perceptually important artifacts. Given the nature of GAN-based solutions, we focus on the sensitivity of human vision to hallucination in case of input samples with different densities. We propose psychophysical experiments, a dataset, and a procedure for training foveated image reconstruction. The proposed strategy renders the generator network flexible by penalizing only perceptually important deviations in the output. As a result, the method emphasized the recovery of perceptually important image features. We evaluated our strategy and compared it with alternative solutions by using a newly trained objective metric, a recent foveated video quality metric, and user experiments. Our evaluations revealed significant improvements in the perceived image reconstruction quality compared with the standard GAN-based training approach.

CCS Concepts: • **Computing methodologies** → **Perception**; Neural networks.

Additional Key Words and Phrases: foveated rendering, image reconstruction, generative adversarial network, perception

1 INTRODUCTION

Wide-field-of-view displays, such as virtual and augmented reality headsets, require efficient methods to generate and transmit high-resolution images. Techniques for reconstructing foveated images seek to solve the problem by leveraging the non-uniform sensitivity of human vision to spatial distortions across a wide field of view and generate high-quality images around only the location of the gaze as indicated by an eye-tracking device. Such foveated systems usually consist of two main steps [24, 50, 58]. First, an image is generated or transmitted in the

Authors' addresses: Luca Surace, luca.surace@usi.ch, Università della Svizzera italiana, Switzerland; Marek Wernikowski, marek.wernikowski@zut.edu.pl, Università della Svizzera italiana, Switzerland and West Pomeranian University of Technology, Poland; Cara Tursun, cara.tursun@rug.nl, Università della Svizzera italiana, Switzerland and University of Groningen, Netherlands; Karol Myszkowski, karol@mpi-inf.mpg.de, Max Planck Institute for Informatics, Germany; Radosław Mantiuk, rmantiuk@wi.zut.edu.pl, West Pomeranian University of Technology, Poland; Piotr Didyk, piotr.didyk@usi.ch, Università della Svizzera italiana, Switzerland.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2023 Copyright held by the owner/author(s).

1544-3558/2023/2-ART

<https://doi.org/10.1145/3583072>

form of a sparse set of samples that are generated according to the location of the gaze. Second, the image is reconstructed from the sparse information before being shown to the observer. An example of such a technique is foveated rendering [15] where fewer image samples are computed for peripheral vision to save computation during rendering (Figure 1).

We focus on the second step above, i.e., reconstructing an image from sparse samples. While simple techniques, such as interpolation [50] can be used to this end, it has been demonstrated that machine-learning techniques, more precisely, generative adversarial networks (GANs), can provide superior results [24] due to their ability to hallucinate missing content based on the learned statistics of the given image or video. Although such reconstruction requires additional computation, it can provide results of a similar quality as those obtained by simpler techniques while using fewer input samples. However, several challenges persist in design and training in this regard. Every GAN architecture is composed of two neural networks trained simultaneously [14, 24]. In the task of foveated image reconstruction, one of them, called the *generator*, is responsible for reconstructing the image from a sparse set of samples, while the second one, called the *discriminator*, is responsible for discriminating between real and reconstructed images. The training iterations of both networks are interleaved such that an improvement in one of them triggers an improvement in the other. Ultimately, training in this manner can be viewed as a game between the generator and the discriminator networks, where the generator tries to reconstruct perfect images from a limited number of samples to try to fool the constantly improving discriminator. Therefore, to successfully train a GAN architecture, maintaining a balance between the training of the generator and the discriminator is highly important. Another challenge, which is the main focus of this paper, involves the choice of training loss and procedure. As in other supervised learning solutions, they have a significant influence on the final performance of GAN-based reconstruction.

The recent literature has acknowledged that for any task where the perceived quality is critical, the loss function must capitalize on visually important image features. A well-known strategy to incorporate such perceptual findings in neural network training is to use a perceptual loss function (e.g., LPIPS [70]). While training a GAN, it is possible to use such a function as the loss for training the generator [24]. However, our main hypothesis is that this is insufficient because the training of the discriminator should also take into account the properties of human visual perception. To address this problem, we propose a new training scheme for the discriminator. Instead of training the discriminator to distinguish reconstructed images from real images, our technique trains it to distinguish the reconstructed images from real images that contain imperceptible distortions.

In this way, the discriminator network can inherit the limitations of the HVS (Human Visual System) represented in the data, and stops penalizing the generator network for reconstructing imperfect images within perceptual limits.

Consequently, in this work, we aim to improve GAN-based machine-learning approaches for foveated reconstruction by introducing a new training scheme based on perceived quality degradation. We first design and conduct psychophysical experiments to study the sensitivity of human visual system to content-based hallucinations across a wide visual field. Our choice of stimuli is inspired by several findings on the degradation of the

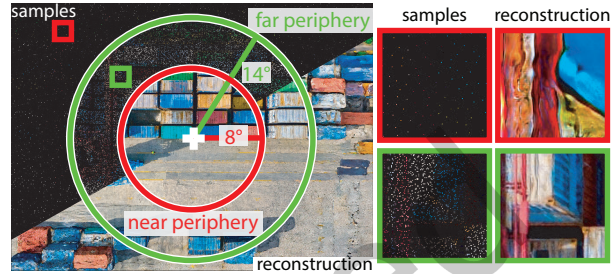


Fig. 1. The input to the foveated image reconstruction are sparse samples (magnify the top-left part of the image or see the insets), based on which the technique reconstructs the images (bottom-right part of the image). The image in this example consists of three regions: fovea (100% samples), near periphery (12% samples), and far periphery (0.7% samples). The white cross in the center indicates the position of the gaze. The reconstructed regions are combined by using linear blending to create a smooth transition between them.

sensitivity of human vision along the periphery. Although many foveated systems have previously exploited the loss of visual sensitivity to high-frequency information [15, 58], the effect does not fully explain the visibility of missing information in peripheral vision. For example, past psychophysical experiments suggest that even though a perfect reconstruction of fine details of the image in the periphery is not critical, a complete lack of high spatial frequencies is detectable [55]. Another effect that is not fully explained by the reduced visual sensitivity to higher spatial frequencies in the periphery is the increased positional uncertainty [20, 31]. To reflect these findings, we employ a technique of texture synthesis guided by the statistics of the original images to generate stimuli with varying amounts of hallucinated content. We argue that this type of distortion resembles the content synthesized by using GAN-based image reconstruction, and our experiments quantify their visibility. We then demonstrate how to incorporate the experimental results into training. Finally, we show how our strategy of focusing on perceptually important image features during training can lead to a GAN-based foveated reconstruction method that provides higher reconstruction quality with the same number of input samples or, conversely, the same perceived quality using fewer samples, leading to savings in bandwidth or rendering time. We argue that this is possible because our foveated reconstruction method aims to recover perceptually important image features that would be otherwise lost due to sub-sampling. The new dataset also allows us to calibrate application-specific objective metrics that predict image quality. We use the new metric and the perceptual experiments to evaluate our new training strategy and compare it with alternative solutions.

2 RELATED WORK

Our work takes inspiration from and bridges the expertise in visual perception, computer graphics, and machine learning. Here, we provide an overview of the relevant works from these fields.

2.1 Foveal vs. peripheral vision

Retina. The perceptual capabilities of the HVS have been extensively studied under different positions of visual stimuli in the visual field. Perception is not uniform across the visual field owing to optical and physiological limitations. Studies on the retina revealed that the density of photoreceptors in the retina is highly heterogeneous [7, 67]. The central region of the retina, called the *fovea centralis* (or *fovea*), is characterized by a relatively high density of cone photoreceptors and retinal ganglion cells (RGCs). This provides foveal vision with a superior perceptual capability compared with non-foveal (or *peripheral*) vision. Although the fovea provides a sharp central vision, it is relatively small and corresponds to approximately 2° of the visual field, which spans up to $160\text{-}170^\circ$ [28]. On the contrary, peripheral vision corresponds to more than 99% of our visual field.

Peripheral contrast sensitivity. To study the differences between foveal and peripheral vision, previous psychophysical studies have focused on measurements of the contrast sensitivity function (CSF), which represents the sensitivity to changes in contrast at different spatial frequencies [3, 26, 32]. Research on the fovea has shown that the human CSF curve has a peak around 4-8 cycles per degree (cpd), with its tail reaching up to 50-60 cpd. Later, Peli *et al.* [38] and, more recently, Chwesiuk *et al.* [6] extended these measurements to peripheral vision, and observed that the decline in contrast sensitivity is characterized by a smaller peak that shifts toward lower spatial frequencies as eccentricity increases. This implies a loss of sensitivity to content with high spatial frequency content in peripheral vision.

Foveated rendering. The differences between foveal and peripheral vision mentioned above have led to gaze-contingent techniques that process and display images depending on the position of the gaze of the observer. Foveated rendering is an actively studied gaze-contingent technique in this domain. It uses the position of the gaze from an eye tracker for a low-resolution image reconstruction in the periphery [5, 15, 25, 34, 37, 50]. These studies have significantly reduced the computational cost of rendering because they reduce the number of pixels

to be rendered [68]. However, their reconstruction methods are mostly based on the simple interpolation of a sub-sampled image and such post-processing steps as temporal antialiasing and contrast enhancement. However, such a simple reconstruction approach does not aim to replace the high-frequency spatial details lost as a result of the undersampling of the underlying content, leading to noticeable degradation in quality.

Hallucinating image details. Psychophysical measurements show that peripheral vision requires a more sophisticated model than a simple boundary between perceptible and imperceptible regions of contrast guided by the shape of the CSF [45, 47]. Thibos *et al.* [55] revealed that the threshold of resolution declines from 14 cpd to 2.6 cpd in the range of eccentricity of 5° to 35° , whereas the threshold of detection drops from 46 cpd to 28 cpd in the same range of eccentricity. As a result of the faster drop-off in the threshold of resolution, there exists a band of spatial frequencies that can be detected but not accurately resolved for each value of eccentricity. Additional studies have shown that performance in terms of discriminating the spatial phase also degrades with increasing eccentricity and leads to greater positional uncertainty in visual perception [31, 35, 40]. Rosenholtz [44] claimed that the HVS encodes image statistics rather than precise location information in peripheral vision, leading to a performance decline in resolving the stimulus position. These studies have important implications for the design of foveated image reconstruction methods because they clearly show that HVS models driving the reconstruction must be comprehensive enough to consider multiple aspects of visual perception. In contrast to the standard reconstruction techniques mentioned above, we address this missing piece in the foveated image reconstruction pipeline.

2.2 Metamers

The goal of foveated rendering can be viewed as the low-cost production of images that are metameric to the full-quality rendering. *Metamer* here refers to images that are structurally different but appear the same to the human observer. Moreover, foveated rendering assumes knowledge about gaze position; therefore, images are metameric usually only for a given gaze location at which most of their content is observed by peripheral vision.

The limitations of the HVS perception have inspired several important studies on metamerism. Initial work aimed to synthesize textural metamers, i.e., different images representing the same type of texture. To this end, Portilla and Simoncelli [39] used an iterative optimization that is run until a randomly initialized image patch converges to the same summary statistics as the target texture. Their observations led to further studies on crowding effects, and Balas *et al.* [2] revealed that the representation of summary statistics can explain the crowding effects observed in the periphery. Rosenholtz *et al.* [46, 47] introduced the texture tiling model (TTM), which models the performance of visual search in the periphery. Based on the ideas on summary statistics, Fridman *et al.* [12] proposed a convolutional network to reproduce the outputs of the TTM, and Deza *et al.* [9] introduced a generative adversarial network model for creating metamers of peripherally viewed natural images. Instead of using hand-crafted summary statistics, as previously introduced by Portilla and Simoncelli [39], they used channel autocorrelation statistics computed from the pre-trained VGG network features [48]. Although these studies have delivered promising results, their main goal is to study foveal texture perception or provide a reference model for studying the properties of peripheral vision (e.g., for visual crowding). More recently, Walton *et al.* [61] proposed a real-time method for producing metameric images to peripherally viewed input images with the main application of image and video compression.

While all the above-mentioned studies have focused on producing a metamer of an input image, this approach is not directly applicable to foveated rendering, the goal of which is to avoid rendering the original, full-quality image in the first place. Therefore, it is interesting to consider the problem of computing images that are metameric to full-quality rendering but are derived based on partial information from the rendering system. Examples of such techniques include standard foveated rendering, where the shading rate is reduced toward the periphery [15], as well as more recent techniques, where contrast enhancement [37] or noise synthesis [53] are applied to

further reduce the amount of information needed for reconstructing peripherally viewed metamers of full-quality images. Kaplanyan *et al.* [24] recently introduced a powerful method to this end. They proposed a foveated image compression solution by using a GAN model that reconstructs perceptually plausible image sequences from a very sparse set of samples while maintaining temporal coherence. Our work takes inspiration from this solution and aims to minimize the required number of samples from the underlying content while achieving the best-perceived quality. To this end, similar to past work, we focus on reconstructing by hallucination but capitalize on positional uncertainty, and distortions [55] by modifying the training scheme. More precisely, in contrast to the scheme presented by Kaplanyan *et al.* [24], where the discriminator network is trained on ground-truth images, we train it on data derived from a series of systematic experiments that analyze the sensitivity of the HVS to distortions in the periphery.

2.3 Image metrics and perceptual loss

One way of guiding image reconstruction is to use image metrics. Current foveal quality metrics use the properties of central vision and provide inaccurate predictions for the periphery. The growing research on peripheral vision and its applications to foveated image reconstruction suggests the need for new foveated metrics [17, 21, 29, 42, 52, 56, 57, 60, 64]. These metrics are promising candidates to guide the loss function in learning-based approaches. However, their complex implementations, costly computations, and, in some cases, non-differentiable operations pose challenges for training models of image reconstruction by using them. An alternative to this is to use a training loss defined on the feature maps from a pre-trained deep network. This has become one of the most common approaches to learning-based image reconstruction, especially for super-resolution-based techniques [10, 23, 70]. Compared with a simpler loss function, such as the mean squared error (MSE), the loss functions defined on the hierarchical features of deep networks more closely resemble how the HVS processes visual information. However, there may still be significant differences between deep network representation and human visual perception [11]. In addition, some of the most commonly used pre-trained networks have been shown to have redundancy in their feature representations when reconstructing for the best-perceived quality [54]. The losses defined on those feature representations improve the perceived quality in the fovea but are not specifically optimized for peripheral vision. In this work, we take an orthogonal approach in the context of GAN training. Apart from using a perceptual loss to train a generator, our main contribution is a modification to the training of the discriminator such that it better reflects the discriminative power of a human observer.

3 PERCEPTUAL EXPERIMENTS

There is an important connection between studies on metamers and those on foveated image reconstructions using GANs discussed above. While the former postulate the importance of preserving image statistics for peripheral vision, the latter reconstruct the content according to the discriminator trained on natural images and videos. However, we argue that training the discriminator using natural images does not adequately reflect the lack of sensitivity of the HVS to spatial distortions, and excessively constrain the generator in hallucinating content. Therefore, we propose to train the discriminator on a dataset composed of images that contain distortions that are unobjectionable to the observers.

It is important to note that the primary goal of this procedure is not to cause the GAN to produce distortions, but rather to make it insensitive to the distortions that humans cannot detect and focus on penalizing perceptually important artifacts. By doing this, we want the discriminator to share limitations similar to those of the HVS.

Training GANs on an extensive dataset of images containing distortions with a near-visibility-threshold requires the responses of human observers in a subjective experiment, in which the participants are asked to adjust the level of distortion in peripherally viewed stimuli. However, owing to the sheer amount of data

typically required for training the GAN, it is unfeasible to generate such a dataset by relying solely on perceptual experiments in a controlled lab environment with a reasonable number of participants. Therefore, we rely on method of texture synthesis that takes advantage of the correlation between image features to preserve the statistical properties of an input image. By imposing pixel-level constraints, we can control how faithful the reconstruction is to the structure of an exemplar. The number of pixel-level constraints acts as a parameter that controls the freedom to change the structure of the synthesized image with respect to the input, thereby introducing a way to control the strength of visual distortions that can be permitted in the reconstruction. We aim to use subjective experiments to measure the strength of distortions, which makes the reconstruction metameric with respect to the peripherally viewed ground-truth stimulus. Once the optimal parameters have been estimated for a smaller set of images in the perceptual experiment, we generate a dataset large enough for training the GAN by using the method of texture synthesis, thus eliminating the need to conduct subjective experiments with an unreasonably large number of participants.

3.1 Stimuli generation

Our stimuli generation model is based on the texture synthesis method proposed by Gatys *et al.* [13]. We customized their method by partially constraining the synthesis to control the level of visual distortion in a convenient way. We exploited this capability to create metamers of the input image given the specific viewing conditions. Their method was formulated as an optimization procedure on feature maps of the pre-trained VGG-19 [48] network that optimizes \hat{x} for an input exemplar \vec{x} by minimizing the loss function:

$$\mathcal{L}(\vec{x}, \hat{x}) = \sum_{l=0}^L w_l \sum_{i,j} \frac{1}{4N_l^2 M_l^2} (\hat{G}_{ij}^l - G_{ij}^l), \quad (1)$$

where \hat{G}_{ij}^l and G_{ij}^l are Gram matrices for feature maps i and j in layer l , N_l is the number of feature maps, M_l is the total number of neurons in a layer, and w_l is an additional weight associated with layer l . To synthesize images for our experiment, we use the same procedure but also constrain a portion of randomly chosen pixel values in \hat{x} such that they are identical to the corresponding pixels in \vec{x} . We refer to these pixels as *guiding samples*. We enforce the quality constraint by projecting the solution to the feasible space in each iteration of a gradient descent optimization.

We observed that solving the constrained optimization leads to subtle image artifacts that resemble checkerboard patterns (Figure 3, first row – middle). The problem is related to the well-known issue of checkerboard artifacts created by backpropagation [36]. We identified two solutions to address this issue. The first consists of running the constrained optimization until convergence, removing the checkerboard artifacts characterized by a high spatial frequency by applying a low-pass Gaussian filter ($\sigma = 1$), and running a second round of optimization without the constraint. We also perceptually verified that similar results may be achieved when constrained optimization is initialized with the guiding samples filtered by a Gaussian filter (Figure 3, second row).

Using the above procedure, we computed the images \hat{x}_p , where $p\%$ is the percentage of the guiding samples (Figure 2, left). For $p = 0$, our synthesis is equivalent to the original technique presented by Gatys *et al.*

The loss of high spatial frequencies (as commonly observed when reducing the input resolution) is an important factor influencing the perceived quality. To improve the sensitivity of the trained metrics to a visible decline in resolution, we also created a separate dataset of images with different degrees of Gaussian blur, i.e., different values of σ of the Gaussian kernel (Figure 2, right). The results of the perceptual experiment obtained with these stimuli were used to expand the dataset of images to train the image metrics in Section 5.2.

3.2 Experimental protocol

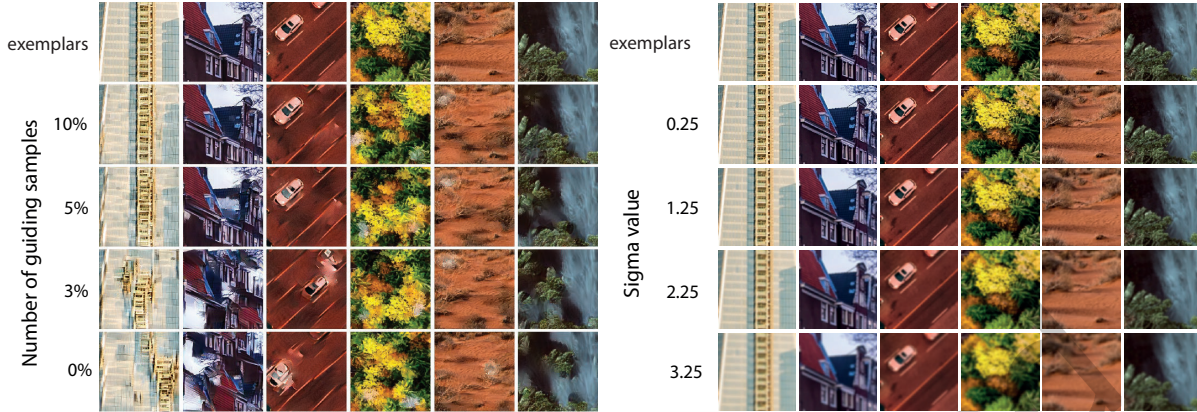


Fig. 2. Results of stimulus generation for different numbers of guiding samples (left) and values of σ (right) of the Gaussian filter. Note the increased distortions relative to the exemplar when the number of guiding samples decreases (left) and when σ increases (right).

The number of the guiding samples p and the value of σ of the Gaussian kernel provide a parametrization for our investigation of the sensitivity of the HVS to deviations relative to the original images. More precisely, with the generation of stimuli using texture synthesis in our main experiment, we sought a direct relationship between the number of guiding samples and the probability of detection of the distortions by a human observer at a particular eccentricity. We later used this relation to generate a much larger dataset of images with imperceptible distortions that is required for training GAN-based foveated reconstruction. By contrast, the additional experiment with Gaussian blur only sought pairs of images and the corresponding probability of blur detection, as a smaller dataset was sufficient for training our image metrics.

Stimuli. We prepared 24 image patches of size 256×256 each from a different 4K image. The images are grouped into two main categories: nature and architecture. Nature images typically do not contain as much structure as the architecture images of human-made objects. The features present in the nature images have a larger variance in their texture, both in terms of colors and frequency-related content. Natural objects have a large variety of shapes without any strict pattern. On the contrary, architecture scenes usually contain larger uniform areas with clear separation between different parts. We expected that reconstructing images with a clear structure may be more challenging. For each patch, we generated a corresponding set of distorted patches for $p \in \{0, 3, 5, 7.5, 10\}$. The set of values of p was determined in a preliminary experiment, in which we found that $p > 10\%$ yields to images that are almost always perceptually indistinguishable from the originals. A set of ground-truth patches was used to generate a set of blurred patches

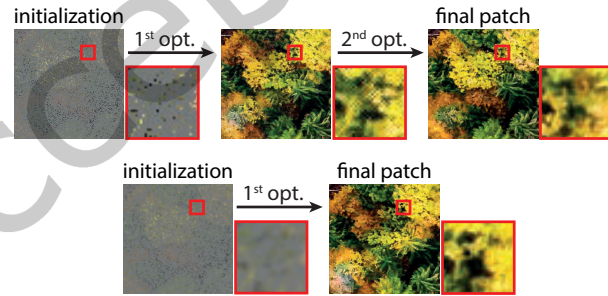


Fig. 3. The figure presents two strategies for synthesizing stimuli by using the guiding samples. Top: a two-step procedure where a second unconstrained optimization is performed to remove subtle checkerboard artifacts after applying a Gaussian filter with $\sigma = 1$. Bottom: constrained optimization initialized with a blurred version of the guiding samples.

for $\sigma \in \{0.25, 1.25, 2.25, 3.25, 4.25\}$. The range was chosen to uniformly span the range of visible blur across the considered field of view [58]. The sample of our stimuli is presented in Figure 2.

Task. The experiment started with a short initial warm-up phase, in which the participants received instructions about the task. Subsequently, in each trial, three patches were presented to them on the screen: (1) the original patch at the fixation point, (2) a synthesized stimulus on either the right or the left side at a given eccentricity, and (3) the original patch on the opposite side at the same eccentricity. The stimuli were visible to both eyes, and the participants were asked to select the patch that was more similar to the reference by pressing left or right arrow keys of the keyboard.

Although we asked the participants to maintain their gaze at the center of the screen during the experiment, involuntary changes in the position of the gaze might have occurred from time to time. To preserve the retinal position of the stimuli against involuntary movements of the eyes of the participants, the stimuli followed the eye movements (Figure 11). For this purpose, the gaze position is continuously monitored by using an eye tracker. The participants were not required to always focus on one point because the stimuli always followed the gaze point. The participants did not receive feedback on the correctness of their responses. We tested the visibility of distortions at 8° (the end of the parafovea [62]), 14° (the center of the perifovea [51, 62]), and 20° , for which the stimuli spanned 3.21° , 3.08° , and 2.89° , respectively. The distance between the participant and the display was set to 70 cm during the experiment. At this viewing distance, each pixel spanned approximately 0.012 visual degrees. We did not impose a limit on the viewing time, and the average duration of each trial was 2 seconds. In total, each participant performed 1800 trials, leading to a total duration of almost 1 hour. The order of the images, eccentricities, and sides on which the test stimuli were shown were randomized. We used these experimental settings with both types of stimuli (created using texture synthesis and Gaussian blur).

Hardware. We used a setup consisting of a 27" Acer Predator display operating at a resolution of 3840×2160 px at 120 Hz and a peak luminance of 170 cd/m^2 . We used a Tobii Pro Spectrum eye tracker at a sampling rate of 600 Hz to track the position of the gaze.

Participants. In order to investigate the potential effects of the participants' background and knowledge of the field, we conducted this experiment¹ with two groups of participants. The first group consisted of five participants (four of them were the authors) who had extensive experience in computer graphics and full knowledge of the task. The second group consisted of 10 naive participants who had no experience in computer graphics or related fields. To improve the diversity of the stimuli, the naive participants performed the experiments with an extended dataset that contained four additional images. All participants were between 25 and 36 years of age and had a normal or corrected-to-normal vision.

3.3 Data analysis

The results of Gaussian blur were used directly to train the image metrics (Section 5.2). The rest of the data, i.e., for texture synthesis-based stimuli, were processed separately for the expert and naive participants, and then compared. For each eccentricity and number of guiding samples tested, we first aggregated the participants' answers across all images and then computed the probability of their detection of distortions in the patches. We then expressed the relationship between the number of guiding samples and the probability of detection for each eccentricity by using a cubic polynomial fit. The results are shown in Figure 4 for both the expert and naive groups.

To construct a set of image patches to train a discriminator in a GAN architecture, we sought a relation between eccentricity and the number of guide samples, to produce patches that did not contain objectionable artifacts when used with texture synthesis. We used the probabilities obtained from the experiments for guidance.

¹approved by Ethical Committee of Università della Svizzera italiana, decision CE.2020.3.

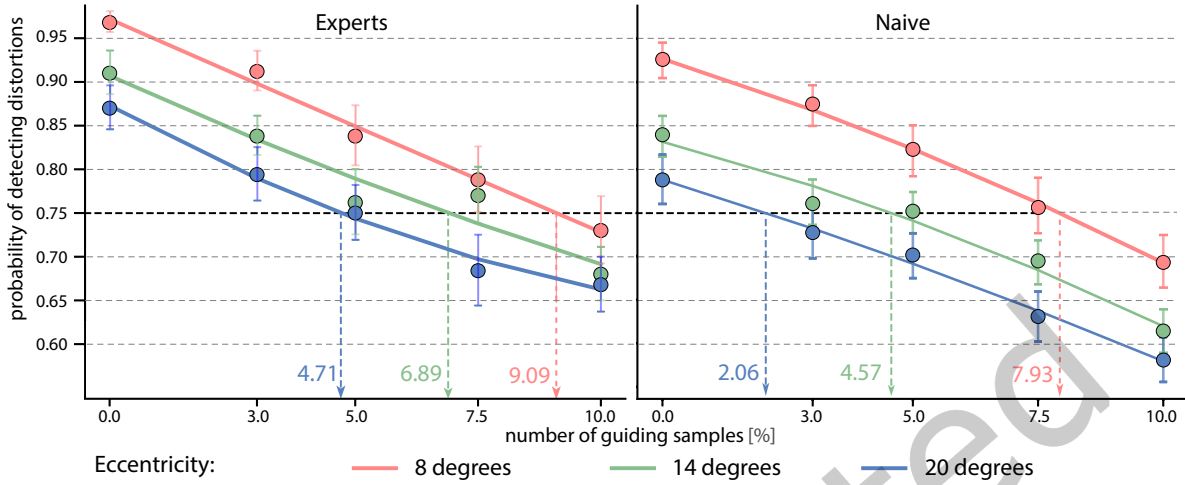


Fig. 4. The probability of detecting differences between the original and the synthesized images as a function of the number of guiding samples and eccentricity. The error bars visualize standard errors of the mean. On the left are the results of the expert group and on the right are those of the naive group.

More precisely, we used the number of the guiding samples corresponding to a 75% probability. Our choice was motivated by the commonly used definition of one just-noticeable-difference (JND) being a transition between visible and invisible distortions [33]. In practice, this is the midpoint between distortions that are always visible and those that are invisible. Although lower probabilities, such as 50%, can be considered, we argue that 75% provides a good trade-off for two reasons. First, the probabilities estimated from psychological experiments can asymptotically converge to only 50%, posing challenges when seeking the exact 50% point. The estimation of the threshold becomes ill-conditioned when the psychometric slope approaches zero, while 75% is an adequately located value as it lies in the steepest part of the psychometric function. Second, our experiments applied an isolated scenario in which the participants were given the particular task of determining the higher-quality patch, and only the reference and the test patch were shown to them. In ultimate applications such as foveated rendering, the observer is less sensitive to any distortion. Therefore, we believe that choosing the number of guiding samples leading to near-threshold distortions was appropriate in this case.

We used a 75% probability as a guide together with cubic fits to our data. We computed the number of guiding samples for expert subjects as 9.09%, 6.89%, and 4.71% with a 95% CI [7.85 - 10.48], [5.78 - 8.14], [3.60 - 5.94] for eccentricities 8°, 14° and 20°, respectively. For the naive participants, the results are: 7.93%, 4.57% and 2.06% with a 95% CI [7.47 - 8.41], [3.98 - 5.29], [1.30 - 2.76]. The confidence intervals were computed by bootstrapping, and the estimates are shown in Figure 4. As expected, the naive observers were less sensitive to the artifacts than the experts, and tolerated distortions in synthesized patches with fewer guiding samples. We used the estimated values to prepare the inputs for the discriminator during GAN training (Section 4.1).

4 METHOD

The results of the perceptual experiments described in the previous section provide the measured thresholds for structural distortions for a standard observer. We used these data to control the learned manifold of the target images in foveated image reconstruction. To this end, we present an improved training scheme for the GAN in which the training data consist of a set of natural and synthesized images.

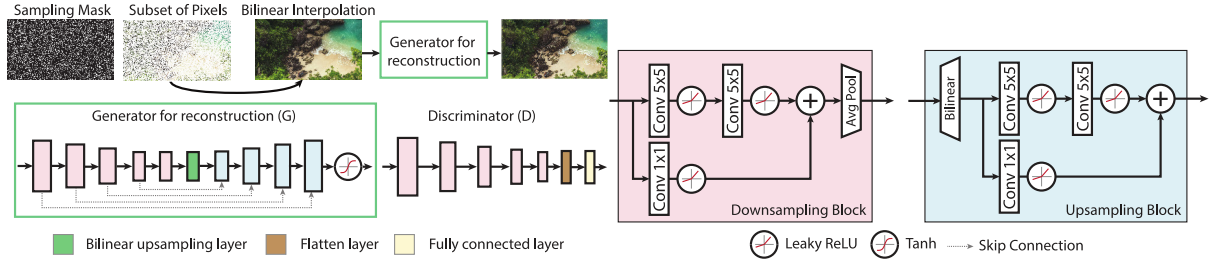


Fig. 5. The architecture of our proposed network.

Our network for the foveated reconstruction uses the Wasserstein GAN [1] training scheme to produce perceptually optimized reconstructions from subsampled images. The network topology is based on the UNet encoder-decoder structure with skip connections [43] (Figure 5). This network design is similar to the model previously used by Kaplanyan *et al.* [24]. The encoder part of the generator network (G) consists of downsampling residual blocks that use average pooling layers [18]. Each residual block of the encoder consists of two convolutional layers with a filter size of 5×5 , except for the main branch, where we use a 1×1 filter to adjust the dimensionality. The numbers of filters are 16-32-64-128-128 in each block, respectively. The decoder part is a mirrored version of the first four encoder blocks with upsampling blocks that use bilinear interpolation instead of average pooling. The encoder and the decoder are connected by an additional bilinear upsampling layer. We use LeakyReLU activation with a negative slope coefficient of $\alpha = 0.2$ throughout the network, except for the final layer of the generator, which uses tanh activation.

The discriminator network or *critic* (D) is based on PatchGAN [22], with a patch size of 64×64 . The discriminator consists of downsampling blocks similar to the encoder part of the generator (the number of filters: 16-32-64-128-128). The output of the downsampling blocks is flattened and passed to a fully connected layer that produces a scalar. Compared with the model developed by Kaplanyan *et al.*, we use a more compact generator with half the number of filters in the first and last three blocks. Such a compact generator is made possible by our more permissive training scheme, which allows for imperceptible deviations from the statistics of the target image. By contrast, the discriminator loss used by Kaplanyan *et al.* aims to match the statistics of the target image as closely as possible. This important difference in our training scheme makes it possible to retain the perceptual quality of the image with a more compact network. Furthermore, their technique aims to reconstruct the original image, while our work aims to show the feasibility of using a GAN trained on perceptual data.

4.1 Dataset

We used two separate datasets as inputs to the generator and the discriminator. The input dataset for the generator consisted of patches from natural images with a size of 256×256 . The patches were generated by cropping images with random offsets. To maintain a balanced data representation, 50 images were randomly selected from each of the 1000 classes in the ImageNet dataset [8], which provided us with a total of 50K patches. These images were later sub-sampled using the void-and-cluster algorithm [59] with sampling rates of 12% for the near periphery and 0.7% for the far periphery. This choice was made according to a content-aware foveated rendering method proposed by Tursun *et al.* [58]. The sub-sampling was followed by bilinear interpolation before the images were passed on to the generator as input.

We used the texture synthesis method described in Section 3.1 using the results of the perceptual experiment described in Section 3.2. This dataset consisted of 50K patches that were synthesized using 9.09% and 6.89% of

the pixels as guiding samples, respectively, for near and far peripheral regions in addition to the full-resolution ground truth images.

4.2 Discriminator loss

The training of the discriminator, D , uses the same loss function as in the original WGAN design [1]:

$$\mathcal{L}_{adv} = D(x) - D(G(z)), \quad (2)$$

where z represents the input to the generator network G , $D(x)$ is the output of the discriminator to real samples (natural images), and $D(G(z))$ is the output of the discriminator to reconstructions from G . This training is equivalent to the optimizations performed in previous work when $x \in \mathbb{I}$, where \mathbb{I} is the set of images from ImageNet.

In our experiments, we updated this formulation by using $x^* \in \mathbb{I}^*$, where \mathbb{I}^* is the set of images with visually imperceptible structural distortions, as we described in Section 3.2. We denote the discriminator loss operating on this manifold of synthesized images with structural distortions by:

$$\mathcal{L}_{adv}^* = D(x^*) - D(G(z)). \quad (3)$$

To ensure the stability of the training of the WGAN, we imposed a soft Lipschitz constraint by using a gradient penalty [16].

4.3 Generator loss

Our optimization trained generator networks with a weighted sum of different types of losses. For a comprehensive evaluation of our training scheme, we focused our analysis on three types of generators trained with standard and perceptual losses. The first generator, G_{L2} , was trained with a combination of standard MSE loss and adversarial loss. The second generator, G_{LPIS} , was trained with the learned perceptual image patch similarity loss term. We used the learned linear weights on top of the VGG network as provided by the authors in their work [70]. Moreover, inspired by [19], we added the generator G_{Lapl} that used Laplacian-based loss. It is defined as a weighted sum of the mean squared error between the corresponding levels of the Laplacian pyramid for the reconstruction and the ground truth. We assigned weights to each level according to a Gaussian with $\sigma = 1.0$. By centering the Gaussian around different levels, we were able to place more emphasis on the reconstruction fidelity of different spatial frequencies in the pyramid decomposition. The main motivation behind the loss was that by assigning a larger weight at lower spatial frequencies, the network will be given more freedom to hallucinate high spatial frequencies, which might be desirable in the periphery.

All the generator losses used in our experiments are listed in Table 1.

4.4 Training

Inspired by work by Guenter *et al.* [15], we considered a foveated rendering scenario in which the image was divided into three regions with different levels of distortions. We assumed that the boundaries of these three regions were at the 8° and 14° of eccentricity (represented as the red and green circles, respectively, in Figure 1), which coincide with the end of the parafovea and its center of periphery [62]. They divided the image into foveal, near periphery, and far periphery regions. Our split was similar to that used in the computer graphics research, such as by [Guenter et al. 2012, Patney et al. 2016].

While the content of the foveal region was directly transferred from a full-resolution image to ensure the highest quality, we reconstructed the near and far periphery regions from a sparse set of samples. The input sampling density for these regions was assumed to be the same as the number of guiding samples estimated in our experiment with image patches (Section 3). More precisely, we assumed that the near periphery region



Fig. 6. Training with (left) and without (right) an input sampling mask.

was reconstructed from the number of samples required for an eccentricity of 8° , and the far periphery was reconstructed from the number of samples corresponding to an eccentricity of 14° .

To perform this reconstruction, we trained two distinct generator networks, each of which was responsible for the reconstruction of the near and far peripheral regions. For benchmarking purposes, we also trained separate networks for each of the two discriminator losses (using our \mathcal{L}_{adv}^* and the standard loss \mathcal{L}_{adv}) and three generator losses (using \mathcal{L}_G^{L2} , \mathcal{L}_G^{LPIPS} , and \mathcal{L}_G^{Lapl}). The relative weights of the loss terms were set to $w_{L2} = 2000$, $w_{LPIPS} = 100$, $w_{Lapl} = 100$, and $w_{adv} = 1$. The selected weights were adjusted according to the magnitudes of the individual loss terms to equalize their contributions to the final loss. This was done by observing the values of the individual terms during training.

We used the Adam optimizer with a learning rate of 2×10^{-5} ($\beta_1 = 0.5$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$). The training lasted for 20–30 epochs until convergence, which took approximately one day on an Nvidia 2080 Ti GPU. We assumed convergence when the training loss reached a plateau. The sample reconstructions from the converged network were also visually checked against potential instabilities during training.

4.5 Sampling mask

By capitalizing on the potential correlation between subsequent frames, the network introduced by Kaplanyan *et al.* uses recurrent connections as an important part of their design to retain information from previously subsampled frames. This high-level temporal reprojection provides the network with additional information when the underlying content is only partially observed owing to sparse subsampling. In order to clearly observe the effects of different training schemes, we used information from only one frame and isolated the reconstruction from the effects of this flow of temporal information. In our initial experiments, we observed that such a design decision made the network more sensitive to the sampling mask used in the inputs due to the absence of the flow of temporal information, which would otherwise have compensated for the lack of information on the true values of the missing samples. In order to address this issue, as a first attempt, we filled in the missing information by interpolating the sampled pixels while retaining the sampling mask as a channel of the input. However, visual inspection revealed visual artifacts collocated with the sampled pixels (Figure 6), and their visibility was dependent on the weights assigned to the loss terms. The effect was the most pronounced when we used \mathcal{L}_{L2} in training, and the artifacts were less visible with \mathcal{L}_{LPIPS} . As a remedy, we removed the sampling mask from the training input and provided the generator with the bilinearly interpolated input consisting of RGB channels, as shown in Figure 5. This solution seemed to be effective in removing visual artifacts from the reconstruction (please refer to Figure 6 for a visual comparison).

Table 1. The loss functions used for training the generator in our evaluations.

Loss function	Definition
L2	$\mathcal{L}_G^{L2} = w_{L2} \cdot \mathcal{L}_{L2} + w_{adv} \cdot \mathcal{L}_{adv}$
L2 ours	$\mathcal{L}_{G^*}^{L2} = w_{L2} \cdot \mathcal{L}_{L2} + w_{adv} \cdot \mathcal{L}_{adv}^*$
LPIPS	$\mathcal{L}_G^{LPIPS} = w_{LPIPS} \cdot \mathcal{L}_{LPIPS} + w_{adv} \cdot \mathcal{L}_{adv}$
LPIPS ours	$\mathcal{L}_{G^*}^{LPIPS} = w_{LPIPS} \cdot \mathcal{L}_{LPIPS} + w_{adv} \cdot \mathcal{L}_{adv}^*$
Laplacian	$\mathcal{L}_G^{Lapl} = w_{Lapl} \cdot \mathcal{L}_{Lapl} + w_{adv} \cdot \mathcal{L}_{adv}$
Laplacian ours	$\mathcal{L}_{G^*}^{Lapl} = w_{Lapl} \cdot \mathcal{L}_{Lapl} + w_{adv} \cdot \mathcal{L}_{adv}^*$

5 RESULTS AND EVALUATION

We evaluated our strategy for training foveated image reconstruction using objective image metrics (Section 5.2) and a subjective experiment (Section 5.3). In our evaluation, we aimed to show that the benefits of our method are not limited to a particular selection of the training loss. To this end, we evaluated the generator network G of our method with six loss functions that were combinations of LPIPS (\mathcal{L}_{LPIPS}), L2 (\mathcal{L}_{L2}), and Laplacian pyramid loss (\mathcal{L}_{Lapl}) terms (Table 1). For the discriminator network, D , we benchmarked the performance of networks trained using our new patch dataset (\mathcal{L}_{adv}^*) as well as the original dataset (\mathcal{L}_{adv}).

5.1 Visual inspection

Figure 7 presents the results of reconstruction obtained by using differently trained architectures on four images. For reference, we include the original high-resolution and the standard foveated reconstruction by using an interpolation with Gaussian weights. For the results of training using Laplacian loss, we introduced a notation consisting of two letters, \mathcal{L}_G^{LaplXY} , where $X, Y \in \{H, M\}$ encode the position of the Gaussian peak at the far and near periphery, respectively. The letter H represents the position of the peak located at the first level of the pyramid (characterized by an emphasis on high spatial frequencies), whereas the letter M represents the position of the peak located at the fourth level of the pyramid (medium spatial frequencies). For example, the method denoted by \mathcal{L}_G^{LaplHM} refers to a reconstruction obtained by using a network trained with Laplacian pyramid-based loss. In this case high spatial frequencies were assigned larger weights for the far periphery, and medium frequencies were given higher importance for the near periphery.

The first observation is that the results of all eight GAN-based reconstruction exhibit clear hallucinated results, and the reconstruction of very fine details is not exact. Although this is visible with direct visual inspection, such deviations are less visible when shown in the periphery. Furthermore, all reconstructions introduce high spatial frequencies and strong edges, but training with \mathcal{L}_G^{L2} loss makes them sparser and more exaggerated. A visual comparison (Figure 7) between the discriminators trained with and without our synthesized dataset (i.e., \mathcal{L}_G^{L2} vs. $\mathcal{L}_{G^*}^{L2}$, \mathcal{L}_G^{LPIPS} vs. $\mathcal{L}_{G^*}^{LPIPS}$, \mathcal{L}_G^{LaplMM} vs. $\mathcal{L}_{G^*}^{LaplMM}$, \mathcal{L}_G^{LaplHH} vs. $\mathcal{L}_{G^*}^{LaplHH}$) shows that our results include higher spatial frequencies. We argue that this is due to the flexibility of the discriminator, which penalizes hallucinations of high spatial frequencies less harshly. This is the desired effect because while the HVS is sensitive to the removal of some high spatial frequencies in the periphery, it is less sensitive to changes in their positions (Section 2).

To further investigate the characteristics of spatial frequency distribution of our reconstructions, we visualized the outputs of frequency band decomposition of the Laplacian pyramid and computed the differences of two layers from the bottom of the pyramid², which encode the highest frequency band as well as one octave below

²Please note the difference between frequency decomposition using the Laplacian pyramid and the Laplacian pyramid-based loss function.

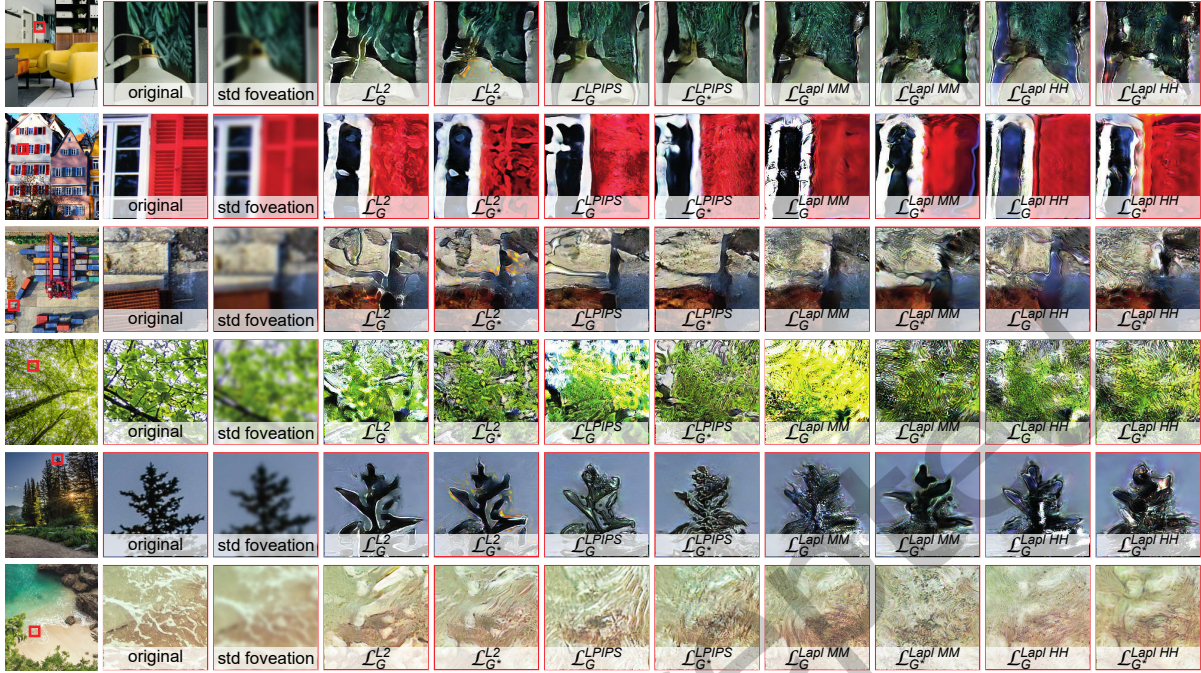


Fig. 7. Sample reconstructions of nature and architecture images by all evaluated methods for the far periphery. $\mathcal{L}_G^{Lapl MM}$ refers to the largest weight assigned to medium spatial frequencies, while $\mathcal{L}_G^{Lapl HH}$ refers to the assignment of the largest weight to high spatial frequencies. Note that the provided images are shown for the demonstration purposes only. For appropriate perception cues, the images corresponding to the figure need to be observed in full screen with 36 cpd. They correspond only to the far peripheral area of vision, spanning from 21° to the corners of the display.

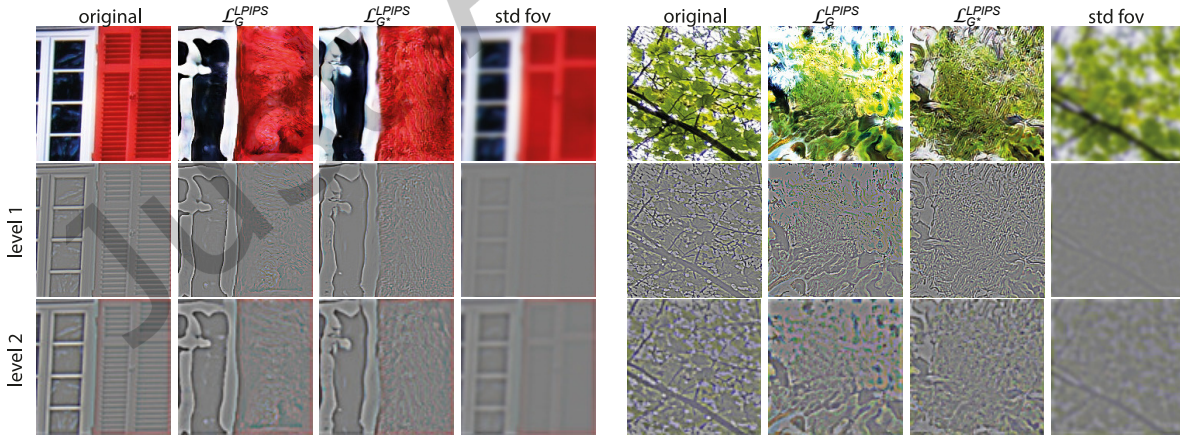


Fig. 8. High-frequency content hallucination using \mathcal{L}_G^{LPIPS} , \mathcal{L}_G^{LPIPS*} , and standard Gaussian blur.

it (Figure 8). We observe that our reconstructions provide additional hallucinated high-frequency details that do not exist in the traditional foveated image reconstruction. Please refer to the supplementary material for an interactive demo with more results.

5.2 Objective image metrics

We assessed the perceptual quality of our foveated image reconstruction using the recently introduced FovVideoVDP metric [33]. FovVideoVDP is a full-reference quality metric that can be used on images and videos. It considers into account the peripheral acuity of the HVS and the retinal eccentricity of the stimuli while computing quality scores. FovVideoVDP quality scores are in Just-Objectable-Difference (JOD) units ($JOD \in [0, 10]$) where $JOD = 10$ represents the highest quality, while lower values represent higher perceived distortion with respect to the reference. We computed FovVideoVDP quality scores of the images generated by our method (\mathcal{L}_{adv}^*) and those reconstructed by networks trained on a standard dataset (\mathcal{L}_{adv}). We provided the original image to the metric as a reference image. We report the FovVideoVDP quality scores in Table 2 for different peripheral regions (near and far), and generator loss functions. We compared these scores with those of our training method using \mathcal{L}_{adv}^* and the standard training approach with \mathcal{L}_{adv} . Our method achieved higher quality-related scores than the standard approach to training the GAN. The generator was able to reconstruct the images better when we included perceptually non-objectionable distortions in the training set of the discriminator using our method.

Table 2. FovVideoVDP [33] quality scores (in JOD units) of our method and standard training of GANs for image reconstruction. The scores were computed for near and far peripheral regions, which represent the images reconstructed at 8° and 14° , respectively. A higher score implies better visual quality.

Peripheral region	Loss function	Ours	Standard
Near (8°)	\mathcal{L}_G^{L2}	8.25	8.08
	\mathcal{L}_G^{LPIPS}	8.19	8.11
Far (14°)	\mathcal{L}_G^{L2}	6.44	6.28
	\mathcal{L}_G^{LPIPS}	6.31	6.18

We also evaluated our method by using other objective quality metrics. Although many objective quality metrics are available for non-foveated quality measurement, objective quality assessment for foveated images is still an open research problem. In the absence of quality metrics for specific types of image distortions, past work has shown that the task-specific calibration of currently available objective quality metrics may be a promising solution [27, 69]. Motivated by this, we used our perceptual data to calibrate existing metrics: L2, SSIM [65], MS-SSIM [66], and LPIPS [70], separately for different eccentricities. The calibration was performed by fitting the following logistic function [41]:

$$y(t) = a + (k - a)/(c + q \cdot e^{-b \cdot t})^{\frac{1}{v}}. \quad (4)$$

to reflect the non-linear relation between the magnitude of distortion in the image and the probability of detecting it, with a, b, c, k, q, v being free parameters. Inspired by LPIPS [70], we also considered reweighing the contributions of each convolution and pooling layer of VGG-19 for each eccentricity separately. We refer to this metric based on the calibrated VGG network as Cal. VGG.

For all metrics, the free parameters (i.e., the parameters of the logistic functions as well as the weights and bias of VGG-19 layers) were obtained by minimizing the mean squared error in predicting the probability of detection:

$$\sum_{(\vec{x}, \hat{x}) \in S_r} \left\| y(M(\vec{x}, \hat{x})) - P(\vec{x}, \hat{x}) \right\|^2, \quad (5)$$

where M is one of the original metrics, S_r is the set of distorted and undistorted pairs of images for eccentricity $r \in 8, 14, 20$, and P is the probability of detecting the difference. Minimization was performed by using nonlinear curve fitting through the *trust-region-reflective* and the *Levenberg-Marquardt* optimizations [4, 30] with multiple random initializations. Furthermore, we constrained the VGG weights to be non-negative values to maintain the positive correlation between image dissimilarity and the magnitude of differences in VGG features, as motivated by the work in [70]. To make our dataset more comprehensive, we added stimuli from an additional experiment that analyzed the visibility of the blur. For this purpose, we followed the procedure described in Section 3.2.

To validate our calibration, we performed 5-fold cross-validation and computed Pearson’s correlations between the ground-truth probability of detecting distortions and metric predictions. Figure 9 presents correlation coefficients for all trained metrics and eccentricities computed as an average across all the folds. Each bar shows the measured correlation for the uncalibrated (bright part) and calibrated (dark part) metrics by using the data from our initial experiment (Section 3.2). For uncalibrated metrics, we used the standard sigmoid logistic function: $y(t) = 1/(1 + e^{-t})$. We also provide the aggregated results, where the correlation was analyzed across all eccentricities. The individual scores showed that as the eccentricity increase, there is a decline in performance in terms of the original metrics. The additional calibration significantly improves the prediction performance in terms of all metrics. An interesting observation is that LPIPS performs very well for small eccentricities (8°). For larger ones (14° and 20°), however, its performance is significantly reduced even with the optimized logistic function. We relate this observation to the fact that LPIPS is not trained for peripheral vision. However, when the weights of the deep layers of VGG-19 are optimized (Cal. VGG), the performance improved significantly. This suggests that the above metrics are promising, but depending on the eccentricity, the contributions of the individual layers to the overall prediction must change. Since our Cal. VGG delivered the best performance in the tests, we selected it to benchmark the foveated image reconstruction techniques listed in Table 1. The results of this test for other metrics that we did not use for evaluation are also reported in the supplementary material as a reference.

After calibration, Cal. VGG is still limited to processing image patches as input. To be able to run Cal. VGG on full images that cover a larger field of view, it needs to consider the influence of changes in eccentricity depending on the position of a given pixel in the image. To support arbitrary values of eccentricity as input, we linearly interpolated the prediction of the metric from 8 and 20 degrees to intermediate values of the eccentricity between 8-20 degrees. Moreover, in contrast to our approach in the calibration step, we switched to a single logistic function whose parameters were estimated by using experimental data from all eccentricities. After these extensions, we ran Cal. VGG locally on non-overlapping patches of the full input image.

To compute a single scalar for the entire image, we took the average value computed across all patches as a global pooling step. To benchmark different reconstruction methods, we randomly selected 10 publicly available images

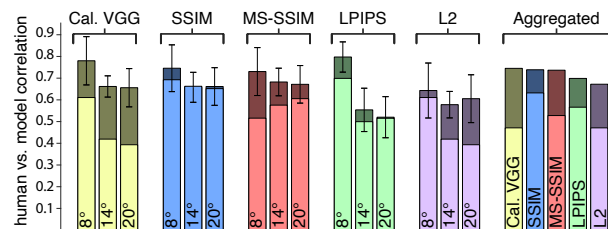


Fig. 9. Pearson’s correlation coefficient of the analyzed methods. Bright bars: uncalibrated metrics (using standard logistic sigmoid with the equation: $y(t) = 1/(1 + e^{-t})$), dark bars: calibrated metrics (fitted to the logistic model using Eq. 4).

at 3840×2160 resolution that contained architectural and natural features. Before applying different reconstruction techniques, we split the images into three regions: fovea, near periphery, and far periphery. We then draw sparse samples as visualized in (Figure 1). To test the reconstruction quality provided by different sampling rates used in near- and far-peripheral regions, we analyzed the Cal. VGG predictions for various blending strategies by changing the eccentricity thresholds at which the transition from near to far peripheral regions occurred. We computed the predicted detection rates from Cal. VGG for threshold points between 9° - 22° for all images. Figure 10 presents the results. Lower detection rates indicate lower probabilities of detecting reconstruction artifacts by human observers and, therefore, a higher reconstruction quality. Training the reconstruction using $\mathcal{L}_{G^*}^{LPIPS}$ and \mathcal{L}_G^{LPIPS} yields reconstructions that are the least likely to be distinguished from the original images. The results generated by using $\mathcal{L}_{G^*}^{L2}$ delivered a lower detection rate than those generated by using \mathcal{L}_G^{L2} . The detection rate for our method is significantly lower when the far periphery threshold is selected in the range of eccentricity of 12° - 22° ($p < 0.05$). For $\mathcal{L}_{G^*}^{LPIPS}$, this difference in detection rate is significant, compared with that for \mathcal{L}_G^{LPIPS} for thresholds between 9° - 16° ($p < 0.05$). We did not note a significant difference between the methods ($p > 0.10$ for all cases considered) when the network was trained by using Laplacian loss. All p -values were computed by using t -test.

We separated the images into two groups according to their prominent visual features: nature and architecture. Nature images were considered to form a class containing fewer geometrical structures and more texture-like areas, e.g., leaves, trees, waves, etc. They usually have a large variety of shapes without any well-defined patterns. In addition, they exhibit a high level of variance in colors and structure. Visual distortions in nature images would be less likely to result in perceivable changes because the variance in color and structure may have a masking effect on the distortions. On the contrary, architecture images mostly contain structures, like human-made objects, such as buildings, and larger uniform areas with clear visual boundaries. They usually have many edges and corners, which makes it more challenging to have a perceptually plausible and faithful reconstruction from sparse image samples. Distorting such images is more likely to lead to the mixing of visual information from different areas, and this is easy to detect even in the peripheral region of vision. Owing to these distinct properties of nature and architecture images, we separately evaluated the results on these two types of images separately. The results show that the difference of detection rate between our method and the standard training, compared to the overall trend, is more pronounced for nature images and less pronounced for architecture images. The results are available in Figure 6 of the supplementary material.

5.3 Subjective experiments

The psychovisual experiment used to derive the data for training our reconstruction methods was performed involved 5 participants. Although it is common to use few participants in such experiments due to their complexity, and given that they should capture the general properties of the HVS, such experiments do not investigate potential differences within the population. In addition, it is not clear whether the method derived from the perceptual data is effective. Therefore, to further validate our claims regarding the new training strategy, and verify the importance of the improvements observed when using calibrated metrics, we conducted an additional subjective user experiment, in which naive participants were asked to directly compare different reconstruction methods.

Stimuli and task. We used the 10 images that were used in our evaluation on objective image metrics (Section 5.2). They were sub-sampled and reconstructed by using $\mathcal{L}_{G^*}^{L2}$, \mathcal{L}_G^{L2} , $\mathcal{L}_{G^*}^{LPIPS}$, and \mathcal{L}_G^{LPIPS} as shown in Figure 1. In each trial, the participants were shown the original image on the left and one of two reconstructions on the right half of the display. The two halves were separated by a 96-pixel-wide gray stripe. The participants could freely switch between reconstructions by using a keyboard. They were asked to select the reconstruction that was more similar to the reference on the left by pressing a key. During the experiment, the images followed the eye movements of the participant, as shown in Figure 11. In contrast to the calibration experiment performed in Section 3.2, in this

experiment we showed full images to the participants, each covering half of the screen. Fixation was enforced as in the calibration experiment described in Section 3.2. Each trial took 15 seconds on average. The total duration of the experiment was around 15 minutes.

Participants. 15 participants with normal or corrected-to-normal vision took part in the experiment. All were naive for the purposes of the study and were given instructions at the beginning. Each participant was asked to compare all pairs of techniques for each image (60 comparisons per participant).

Results. To analyze the results, for each pair of techniques, we computed a preference rate of method A over method B. The rate expresses the percentage of trials in which method A was chosen as visually more similar to the original image. Table 3 shows the preference rates obtained by using networks trained using our procedure ($\mathcal{L}_{G^*}^{LPIPS}$, $\mathcal{L}_{G^*}^{L2}$) in comparison with the standard procedure (\mathcal{L}_G^{LPIPS} , \mathcal{L}_G^{L2}). We report the results for all images (last column) and do so separately for nature and architecture images. We used the binomial test to compute the p -values. The reconstructions

obtained by using a network trained with our strategy are preferred in 57% of the cases ($p = 0.013$). The difference is significant for $\mathcal{L}_{G^*}^{LPIPS}$ with a 59% preference ($p = 0.04$). In the context of different image classes, our method performed well on nature images, both when we consider $\mathcal{L}_{G^*}^{L2}$ and $\mathcal{L}_{G^*}^{LPIPS}$ separately and when we consider them jointly (for each case preferred in 75% of cases, $p < 0.001$). On architecture images, we observe the preference for $\mathcal{L}_{G^*}^{L2}$ (63%, $p = 0.037$). For all techniques, our method is preferred in 40% of the cases ($p = 0.018$). This is consistent with the results of Cal. VGG (Section 5.2), where our method had a lower probability of detection of nature images and a similar probability of detection of architecture images. We hypothesize that there might be several reasons for its poorer performance on architecture images. First, architecture images contain objects with simple shapes, uniform areas, edges, and corners. Such features might not have been represented well in our calibration, where we used 256×256 patches, whose size was limited to avoid testing the visibility across a wide range of

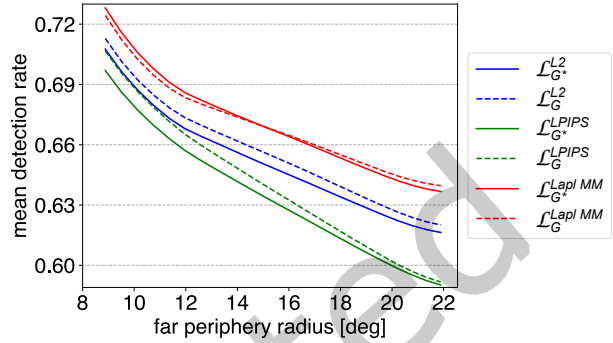


Fig. 10. Detection rates according to the objective Cal. VGG metric with increasing radius of the far periphery. Lower values indicate a higher quality of reconstruction.

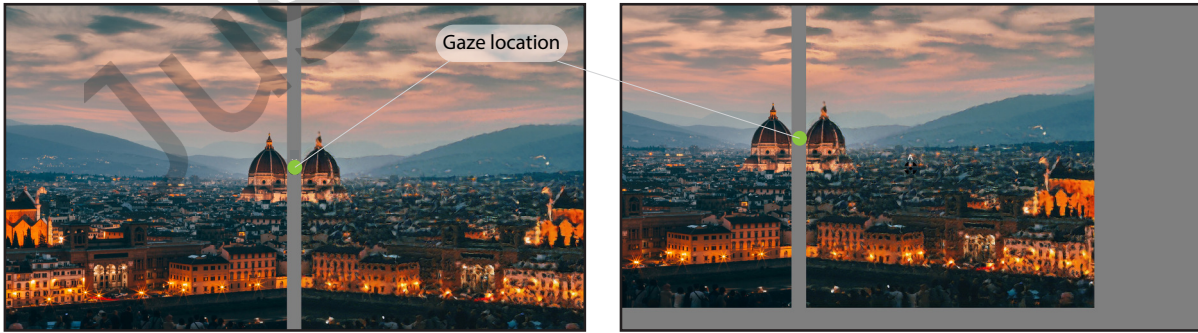


Fig. 11. An example of the stimuli shown during the experiment. The image follows the gaze point, which is marked with the green dot.

Table 3. Preference rates of the methods $\mathcal{L}_{G^*}^{L2}$, $\mathcal{L}_{G^*}^{LPIPS}$ over \mathcal{L}_G^{L2} , \mathcal{L}_G^{LPIPS} when trained using the data collected from the expert group. The values were computed by taking the average across participants. The errors correspond to the standard error of the mean.

	Nature	Architecture	All
$\mathcal{L}_{G^*}^{L2}$ vs \mathcal{L}_G^{L2}	0.75 \mp 0.05	0.37 \mp 0.04	0.56 \mp 0.03
$\mathcal{L}_{G^*}^{LPIPS}$ vs \mathcal{L}_G^{LPIPS}	0.75 \mp 0.05	0.43 \mp 0.07	0.59 \mp 0.05
All ours vs All	0.75 \mp 0.04	0.40 \mp 0.04	0.57 \mp 0.03

eccentricities. Furthermore, we believe that distortions in the visual features of simple objects are much easier to perceive than those in natural textures, which are more random. This problem might have been aggravated because our calibration considered both groups together and did not make any distinction when modeling the perception of artifacts for them. This problem might be solved by using different numbers of guiding samples for different classes of images when generating the dataset for training GAN-based reconstruction. However, this would require more careful data collection for the initial experiment and a more complex model that can predict the number of guiding samples based on the image content. Once these challenges have been addressed, the proposed approach can yield a more accurate dataset and can be used to train a single architecture that can handle different types of images.

Figure 12 shows the preferences for the individual methods compared with those for all other training strategies, including different loss functions, i.e., $\mathcal{L}_{G^*}^{LPIPS}$, \mathcal{L}_G^{LPIPS} , \mathcal{L}_G^{L2} , and $\mathcal{L}_{G^*}^{L2}$. In the experiment with experts (left), $\mathcal{L}_{G^*}^{LPIPS}$ attained the highest preference of 38% ($p < 0.001$) while \mathcal{L}_G^{L2} recorded the lowest preference (24%, $p < 0.001$). When divided into classes, $\mathcal{L}_{G^*}^{LPIPS}$ and \mathcal{L}_G^{L2} were the most preferred methods for natural images, with values of 41% ($p < 0.001$) and 37% ($p = 0.003$), respectively. The other methods had lower preference values - 22% for \mathcal{L}_G^{LPIPS} ($p < 0.001$) and 20% for \mathcal{L}_G^{L2} ($p < 0.001$). \mathcal{L}_G^{LPIPS} was the most preferred on architecture images (37%, $p = 0.005$), followed by $\mathcal{L}_{G^*}^{LPIPS}$ (35%, $p = 0.032$). The $\mathcal{L}_{G^*}^{L2}$ was selected the fewest number of times (21%, $p < 0.001$). All p -values were computed by using the binomial test, and the remaining results were not statistically significant. The experiment, when repeated with naive participants (right), yielded a different threshold as a function of the guiding samples needed for the appropriate foveated reconstruction. In particular, the values related to 8° and 14° changed from 9.09 to 7.93, and from 6.89 to 4.57, respectively. This means that for a standard observer, the number of samples needed to generate an image of fixed quality is higher than the samples needed for an expert observer. Texture synthesis was the initial step of our pipeline. For this reason, we trained all our networks again and we repeated the validation experiment with the new reconstruction. The results are presented in Table 4 and Figure 12 (right). The new experiments showed that while our technique maintained a slight advantage through $\mathcal{L}_{G^*}^{LPIPS}$ over the standard method on nature images, our reconstructions delivered the worst performance on architecture images.

5.4 Discussion

The results of the final evaluation demonstrate that with a successfully derived training dataset containing near-threshold distortions, our GAN-based training strategy can improve the quality of the reconstructions. The benefits, however, are observed with a more conservative calibration (Section 3) performed by experts. When calibration data from the naive participants are used, the final preference shifts towards the standard reconstruction strategy. This shows that our strategy works under conservative calibration conditions. We believe that this owes itself to the potential limitations of our calibration, which was performed on small patches, while larger patches may render some artifacts more prominent. However, there is a trade-off between using small

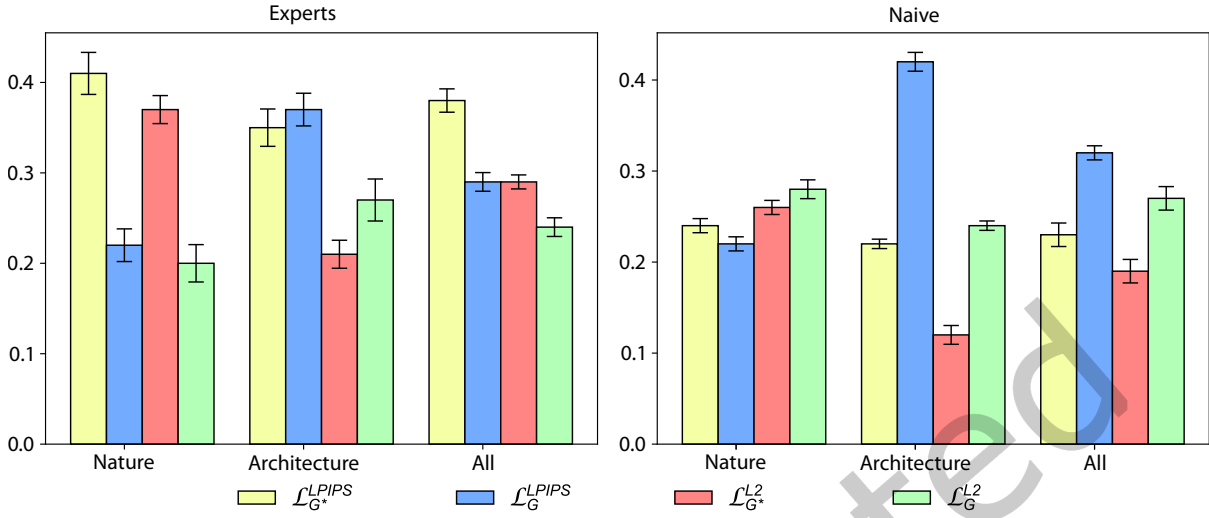


Fig. 12. Preference rates of the methods trained using data gathered from the expert group (left) vs. the naive group (right) in the calibration experiment (Section 3.2). The scores on the y -axis represent the number of times a given method was selected, divided by the total number of trials of the experiment. The error bars show the standard error of the mean. The values were computed by taking an average across participants.

Table 4. Preference rates of the methods $\mathcal{L}_{G^*}^{L2}$, $\mathcal{L}_{G^*}^{LPIPS}$ over \mathcal{L}_G^{L2} , \mathcal{L}_G^{LPIPS} when trained by using the data gathered from the naive group. The values were computed by taking the average across participants. The errors correspond to the standard error of the mean.

	Nature	Architecture	All
$\mathcal{L}_{G^*}^{L2}$ vs \mathcal{L}_G^{L2}	0.38 \pm 0.05	0.32 \pm 0.05	0.35 \pm 0.02
$\mathcal{L}_{G^*}^{LPIPS}$ vs \mathcal{L}_G^{LPIPS}	0.59 \pm 0.04	0.17 \pm 0.06	0.38 \pm 0.04
All ours vs All	0.49 \pm 0.03	0.24 \pm 0.04	0.36 \pm 0.03

patches and obtaining localized information about the sensitivity to distortions for a particular eccentricity, and making the patches larger and losing this property. In future work, it would be interesting to investigate better calibration strategies for our technique.

As described, our technique can be directly applied to train GAN-based foveated image reconstructions. Our experiments demonstrate that the same network architecture with the same input provides higher-visual-quality reconstructions if our training strategy is used than otherwise. This applies directly to techniques such as the one proposed by Kaplanyan et al. [24]. Because the density of the input sampling can be changed, we argue that our training can also reduce the number of input samples while preserving quality, which will improve the efficiency of the entire image generation pipeline.

6 CONCLUSIONS AND FUTURE WORK

Currently available techniques for foveated image reconstruction use perceptual loss to guide network training to capitalize on perceptually important image features. The goal of this work was to inject perceptual information into the discriminator network. To this end, during training, we provided the discriminator with images containing

distortions that are imperceptible to a human observer. This allowed the discriminator to inherit the properties of the HVS encoded in the training dataset. Our new dataset contains images with invisible spatial distortions based on texture synthesis. We argue that such distortions are much closer to artifacts introduced by GAN-based reconstruction than the previously considered blur. Moreover, the new dataset allowed us to train several image metrics to improve their predictions of stimulus quality presented in the periphery.

We studied the suitability of the new training strategy for foveated image reconstruction. In future work, it is essential to extend this investigation to video content as this may yield benefits when the sensitivity of the HVS to temporal artifacts is incorporated. We trained separate networks for the near and far periphery. While this makes the training procedure easier, a more practical solution is to train one network to handle spatially varying density. An alternative solution to attain this goal is to use a fully convolutional network in the log-polar domain [49, 63]. We also did not focus on computational performance. At present, our unoptimized inference takes 3 seconds on our hardware. Although previous work [24] has shown the feasibility of using GAN in such scenarios, computational efficiency remains an important concern. We believe that making networks and their training aware of the limitations of human perception will be important to close the gaps.

Another exciting direction of research is to design a foveated image metric that accounts for a wide range of effects. While work by Mantiuk et al. [33] has taken this approach, they targeted image quality instead of the visibility of distortion. The challenge here lies in collecting large-scale perceptual data with eye-tracking-based information that can help determine the visibility of distortions. Even though our dataset contained this information, it is not sufficient to train a general-purpose visibility metric for both foveal and peripheral vision.

Finally, we believe that the idea of supplying the discriminator in the GAN architectures with images containing near-threshold distortions during training extends beyond applications to foveated image reconstructions. We see it as a more general strategy for training perception-aware GAN-based techniques for the creation of graphical content. Our work considered texture synthesis as a technique suitable for creating controlled distortions relevant to our application. However, it is not ideal for capturing other characteristics of perception, such as the sensitivity to temporal changes, color, and depth, that might be relevant in different applications. It would be interesting if other researchers followed our procedure and included these aspects in the training of generative-adversarial networks to verify their benefits.

ACKNOWLEDGMENTS

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement N^o 804226 PERDY).

REFERENCES

- [1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *arXiv:1701.07875 [cs, stat]* (Jan. 2017). arXiv: 1701.07875.
- [2] Benjamin Balas, Lisa Nakano, and Ruth Rosenholtz. 2009. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of vision* 9, 12 (2009).
- [3] Peter G. J. Barten. 1999. *Contrast sensitivity of the human eye and its effects on image quality*. SPIE press.
- [4] Mary Ann Branch, Thomas F. Coleman, and Yuying Li. 1999. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing* 21, 1 (1999).
- [5] Valentin Bruder, Christoph Schulz, Ruben Bauer, Steffen Frey, Daniel Weiskopf, and Thomas Ertl. 2019. Voronoi-Based Foveated Volume Rendering. In *EuroVis (Short Papers)*. The Eurographics Association, 5.
- [6] Michał Chwesiuk and Radosław Mantiuk. 2019. Measurements of contrast sensitivity for peripheral vision. In *ACM Symposium on Applied Perception 2019 (Barcelona, Spain) (SAP ’19)*. Association for Computing Machinery, New York, NY, USA, Article 20, 9 pages. <https://doi.org/10.1145/3343036.3343123>
- [7] Christine A. Curcio, Kenneth R. Sloan, Robert E. Kalina, and Anita E. Hendrickson. 1990. Human photoreceptor topography. *Journal of comparative neurology* 292, 4 (1990), 497–523.
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*. IEEE, 248–255.

- [9] Arturo Deza, Aditya Jonnalagadda, and Miguel Eckstein. 2017. Towards Metamerism via Foveated Style Transfer. *arXiv:1705.10041 [cs]* (May 2017). arXiv: 1705.10041.
- [10] Alexey Dosovitskiy and Thomas Brox. 2016. Generating images with perceptual similarity metrics based on deep networks. In *Advances in neural information processing systems*. 658–666.
- [11] Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. 2019. Metamers of neural networks reveal divergence from human perceptual systems. In *Advances in Neural Information Processing Systems*. 10078–10089.
- [12] Lex Fridman, Benedikt Jenik, Shaiyan Keshvari, Bryan Reimer, Christoph Zetsche, and Ruth Rosenholtz. 2017. A Fast Foveated Fully Convolutional Network Model for Human Peripheral Vision. *arXiv:1706.04568 [cs]* (2017). arXiv:1706.04568 [cs.NE]
- [13] Leon Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems*. 262–270.
- [14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems 27* (2014).
- [15] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D Graphics. *ACM Transactions on Graphics* 31, 6 (Nov. 2012), 164:1–164:10.
- [16] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved training of Wasserstein GANs. In *Advances in neural information processing systems*. 5767–5777.
- [17] Peiyao Guo, Qiu Shen, Zhan Ma, David J. Brady, and Yao Wang. 2018. Perceptual Quality Assessment of Immersive Images Considering Peripheral Vision Impact. *arXiv:1802.09065 [cs]* (Feb. 2018). arXiv: 1802.09065.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [19] Alexander Hepburn, Valero Laparra, Ryan McConville, and Raul Santos-Rodriguez. 2019. Enforcing perceptual consistency on generative adversarial networks by using the normalised laplacian pyramid distance. *arXiv preprint arXiv:1908.04347* (2019).
- [20] Robert F. Hess and David Field. 1993. Is the increased spatial uncertainty in the normal periphery due to spatial undersampling or uncalibrated disarray? *Vision research* 33, 18 (1993), 2663–2670.
- [21] Chih-Fan Hsu, Anthony Chen, Cheng-Hsin Hsu, Chun-Ying Huang, Chin-Laung Lei, and Kuan-Ta Chen. 2017. Is Foveated Rendering Perceivable in Virtual Reality?: Exploring the Efficiency and Consistency of Quality Assessment Methods. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*. ACM, New York, NY, USA, 55–63. <https://doi.org/10.1145/3123266.3123434> event-place: Mountain View, California, USA.
- [22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.
- [23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and super-Resolution. In *Computer Vision – ECCV 2016 (Lecture Notes in Computer Science)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, 694–711.
- [24] Anton Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. 2019. Deepfovea: neural reconstruction for foveated rendering and video compression using learned natural video statistics. In *ACM SIGGRAPH 2019 Talks (Los Angeles, California) (SIGGRAPH '19)*. ACM, New York, NY, USA, Article 58, 2 pages.
- [25] Jonghyun Kim, Zander Majercik, Peter Shirley, Josef Spjut, Morgan McGuire, David Luebke, Youngmo Jeong, Michael Stengel, Kaan Akşit, Rachel Albert, Ben Boudaoud, Trey Greer, Joohwan Kim, and Ward Lopes. 2019. Foveated AR: dynamically-foveated augmented reality display. *ACM Transactions on Graphics* 38, 4 (July 2019).
- [26] Kil Joong Kim, Rafal Mantiuk, and Kyoung Ho Lee. 2013. Measurements of achromatic and chromatic contrast sensitivity functions for an extended range of adaptation luminance. In *Human vision and electronic imaging XVIII*, Vol. 8651. International Society for Optics and Photonics, 86511A.
- [27] Vamsi Kiran Adhikarla, Marek Vinkler, Denis Sumin, Rafal K. Mantiuk, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. 2017. Towards a Quality Metric for Dense Light Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [28] George F. Koob, Michel Le Moal, and Richard F. Thompson. 2010. *Encyclopedia of behavioral neuroscience*. Elsevier.
- [29] Sanghoon Lee, M. S. Pattichis, and A. C. Bovik. 2002. Foveated video quality assessment. *IEEE Transactions on Multimedia* 4, 1 (March 2002), 129–132. <https://doi.org/10.1109/6046.985561>
- [30] Kenneth Levenberg. 1944. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics* 2, 2 (1944), 164–168.
- [31] Dennis M. Levi, Stanley A. Klein, and Yen Lee Yap. 1987. Positional uncertainty in peripheral and amblyopic vision. *Vision research* 27, 4 (1987), 581–597.
- [32] James Mannos and David Sakrison. 1974. The effects of a visual fidelity criterion of the encoding of images. *IEEE Transactions on Information Theory* 20, 4 (1974), 525–536.

- [33] Rafal K Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. 2021. FovVideoVDP: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics* 40, 4 (2021).
- [34] Xiaoxu Meng, Ruofei Du, Matthias Zwicker, and Amitabh Varshney. 2018. Kernel Foveated Rendering. *Proceedings of ACM on Computer Graphics and Interactive Techniques* 1, 1 (July 2018).
- [35] M. Concetta Morrone, David C. Burr, and Donatella Spinelli. 1989. Discrimination of spatial phase in central and peripheral vision. *Vision research* 29, 4 (1989), 433–445.
- [36] Augustus Odena, Vincent Dumoulin, and Chris Olah. 2016. Deconvolution and checkerboard artifacts.
- [37] Anjul Patney, Marco Salvi, JooHwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. 2016. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics* 35, 6 (Nov. 2016), 179:1–179:12.
- [38] Eli Peli, Jian Yang, and Robert B. Goldstein. 1991. Image invariance with changes in size: The role of peripheral contrast thresholds. *JOSA A* 8, 11 (1991), 1762–1774.
- [39] Javier Portilla and Eero P. Simoncelli. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision* 40, 1 (2000), 49–70.
- [40] Ingo Rentschler and Bernhard Treutwein. 1985. Loss of spatial phase relationships in extrafoveal vision. *Nature* 313, 6000 (1985), 308–310.
- [41] F. J. Richards. 1959. A flexible growth function for empirical use. *Journal of experimental Botany* 10, 2 (1959), 290–301.
- [42] Snježana Rimac-Drlje, Mario Vranješ, and Drago Žagar. 2010. Foveated mean squared error—a novel video quality metric. *Multimedia Tools and Applications* 49, 3 (Sept. 2010), 425–445. <https://doi.org/10.1007/s11042-009-0442-1>
- [43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.
- [44] Ruth Rosenholtz. 2011. What your visual system sees where you are not looking. In *SPIE: Human Vision and Electronic Imaging XVI*, Bernice E. Rogowitz and Thrasyvoulos N. Pappas (Eds.). San Francisco Airport, California, USA. <https://doi.org/10.1117/12.876659>
- [45] Ruth Rosenholtz. 2016. Capabilities and Limitations of Peripheral Vision. *Annual review of vision science* 2 (2016), 437–457. <https://doi.org/10.1146/annurev-vision-082114-035733>
- [46] Ruth Rosenholtz, Jie Huang, and Krista A. Ehinger. 2012. Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in psychology* 3 (2012), 13.
- [47] Ruth Rosenholtz, Jie Huang, Alvin Raj, Benjamin J. Balas, and Livia Ilie. 2012. A summary statistic representation in peripheral vision explains visual search. *Journal of Vision* 12, 4 (April 2012). <https://doi.org/10.1167/12.4.14>
- [48] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*. arXiv:1409.1556 [cs.CV]
- [49] Fabio Solari, Manuela Chessa, and Silvio P. Sabatini. 2012. Design strategies for direct multi-scale and multi-orientation feature extraction in the log-polar domain. *Pattern Recognition Letters* 33, 1 (2012), 41–51.
- [50] Michael Stengel, Steve Grogoric, Martin Eisemann, and Marcus Magnor. 2016. Adaptive Image-Space Sampling for Gaze-Contingent Real-time Rendering. *Computer Graphics Forum* 35, 4 (2016), 129–139.
- [51] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. 2011. Peripheral vision and pattern recognition: A review. *Journal of vision* 11, 5 (2011).
- [52] Nicholas T. Swafford, José A. Iglesias-Guitian, Charalampos Koniaris, Bochang Moon, Darren Cosker, and Kenny Mitchell. 2016. User, Metric, and Computational Evaluation of Foveated Rendering Methods. In *Proceedings of the ACM Symposium on Applied Perception (SAP '16)*. ACM, New York, NY, USA, 7–14. <https://doi.org/10.1145/2931002.2931011>
- [53] Taimoor Tariq, Cara Tursun, and Piotr Didyk. 2022. Noise-Based Enhancement for Foveated Rendering. *ACM Trans. Graph.* 41, 4, Article 143 (jul 2022). <https://doi.org/10.1145/3528223.3530101>
- [54] Taimoor Tariq, Okan Tarhan Tursun, Munchurl Kim, and Piotr Didyk. 2020. Why are deep representations good perceptual quality features?. In *The European Conference on Computer Vision (ECCV)*.
- [55] L. N. Thibos, F. E. Cheney, and D. J. Walsh. 1987. Retinal limits to the detection and resolution of gratings. *JOSA A* 4, 8 (1987), 1524–1529.
- [56] Huyen T. T. Tran, Duc V. Nguyen, Nam Pham Ngoc, Trang H. Hoang, and Truong Cong Thang. 2019. Impacts of Retina-related Zones on Quality Perception of Omnidirectional Image. *arXiv:1908.06239 [cs, eess]* (Aug. 2019). arXiv: 1908.06239.
- [57] W. Tsai and Y. Liu. 2014. Foveation-based image quality assessment. In *2014 IEEE Visual Communications and Image Processing Conference*. 25–28. <https://doi.org/10.1109/VCIP.2014.7051495>
- [58] Okan Tarhan Tursun, Elena Arabadzhyska-Koleva, Marek Wernikowski, Radoslaw Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk. 2019. Luminance-contrast-aware foveated rendering. *ACM Transactions on Graphics* 38, 4 (2019).
- [59] Robert A. Ulichney. 1993. Void-and-cluster method for dither array generation. In *Human Vision, Visual Processing, and Digital Display IV*, Vol. 1913. International Society for Optics and Photonics, 332–343.
- [60] Mario Vranješ, Snježana Rimac-Drlje, and Denis Vranješ. 2018. Foveation-based content adaptive root mean squared error for video quality assessment. *Multimedia Tools and Applications* 77, 16 (Aug. 2018), 21053–21082. <https://doi.org/10.1007/s11042-017-5544-6>

- [61] David R Walton, Rafael Kuffner Dos Anjos, Sebastian Friston, David Swapp, Kaan Akşit, Anthony Steed, and Tobias Ritschel. 2021. Beyond blur: Real-time ventral metamers for foveated rendering. *ACM Transactions on Graphics* 40, 4 (2021).
- [62] Brian Wandell and Stephen Thomas. 1997. Foundations of vision. *Psychocritiques* 42, 7 (1997).
- [63] Panqu Wang and Garrison W. Cottrell. 2017. Central and peripheral vision for scene recognition: A neurocomputational modeling exploration. *Journal of vision* 17, 4 (2017).
- [64] Zhou Wang, Alan C. Bovik, Ligang Lu, and Jack L. Kouloheris. 2001. Foveated wavelet image quality index. In *Applications of Digital Image Processing XXIV*, Andrew G. Tescher (Ed.). San Diego, CA, 42–52. <https://doi.org/10.1117/12.449797>
- [65] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.
- [66] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. Ieee, 1398–1402.
- [67] Andrew B. Watson. 2014. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of Vision* 14, 7 (2014).
- [68] Martin Weier, Michael Stengel, Thorsten Roth, Piotr Didyk, Elmar Eisemann, Martin Eisemann, Steve Grogorick, André Hinkenjann, Ernst Kruijff, Marcus Magnor, et al. 2017. Perception-driven Accelerated Rendering. *Computer Graphics Forum* 36, 2 (2017), 611–643.
- [69] K. Wolski, D. Giunchi, S. Kinuwaki, P. Didyk, K. Myszkowski, A. Steed, and R. K. Mantiuk. 2019. Selecting texture resolution using a task-specific visibility metric. *Computer Graphics Forum* 38, 7 (2019), 685–696. <https://doi.org/10.1111/cgf.13871>
- [70] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 586–595.