Efficient Dataflow Modeling of Peripheral Encoding in the Human Visual System

Efficient Dataflow Modeling of Peripheral Encoding in the Human Visual System

RACHEL BROWN, NVIDIA Research
VASHA DUTELL, UC Berkeley
BRUCE WALTER, Cornell University
RUTH ROSENHOLTZ, Massachusetts Institute of Technology
PETER SHIRLEY, MORGAN MCGUIRE, and DAVID LUEBKE, NVIDIA Research

Computer graphics seeks to deliver compelling images, generated within a computing budget, targeted at a specific display device, and ultimately viewed by an individual user. The foveated nature of human vision offers an opportunity to efficiently allocate computation and compression to appropriate areas of the viewer's visual field, of particular importance with the rise of high-resolution and wide field-of-view display devices. However, while variations in acuity and contrast sensitivity across the field of view have been well-studied and modeled, a more consequential variation concerns peripheral vision's degradation in the face of clutter, known as crowding. Understanding of peripheral crowding has greatly advanced in recent years, in terms of both phenomenology and modeling. Accurately leveraging this knowledge is critical for many applications, as peripheral vision covers a majority of pixels in the image. We advance computational models for peripheral vision aimed toward their eventual use in computer graphics. In particular, researchers have recently developed high-performing models of peripheral crowding, known as "pooling" models, which predict a wide range of phenomena but are computationally inefficient. We reformulate the problem as a dataflow computation, which enables faster processing and operating on larger images. Further, we account for the explicit encoding of "end stopped" features in the image, which was missing from previous methods. We evaluate our model in the context of perception of textures in the periphery, including a novel texture dataset and updated textural descriptors. Our improved computational framework may simplify development and testing of more sophisticated, complete models in more robust and realistic settings relevant to computer graphics.

CCS Concepts: • **Computing methodologies** → **Perception**; *Image manipulation*; *Image compression*; Modeling and simulation;

Additional Key Words and Phrases: Human vision, perception, foveated rendering, image compression

ACM Reference format:
Rachel Brown, Vasha DuTell, Bruce Walter, Ruth Rosenholtz, Peter Shirley, Morgan McGuire, and David Luebke. 2023. Efficient Dataflow Modeling of Peripheral Encoding in the Human Visual System. *ACM Trans. Appl. Percept.* 20, 1, Article 1 (January 2023), 22 pages.
https://doi.org/10.1145/3564605

Authors' addresses: R. Brown, P. Shirley, M. McGuire, and D. Luebke, NVIDIA Research: 2788 San Tomas Expy, Santa Clara, CA 95051; emails: {rabrown, pshirley}@nvidia.com, morgan3d@gmail.com, dluebke@nvidia.com; V. DuTell, UC Berkeley: Minor Hall, University of California, Berkeley, CA 94720-2020; email: vasha@berkeley.edu; B. Walter, Cornell University: 402 Bill & Melinda Gates Hall, 107 Hoy Rd, Ithaca, NY 14850; email: bruce.walter@cornell.edu; R. Rosenholtz, MIT: 32 Vassar St, Cambridge, MA 02139; email: rruth@mit.edu.

1544-3558/2023/01-ART1 $15.00
https://doi.org/10.1145/3564605

1 INTRODUCTION

Computer graphics traditionally renders images with equal detail across the field of view, but researchers have long recognized that a viewer can only perceive a limited amount of this detail at a given moment. The human visual system concentrates sensor resolution, encoding, and processing on the small region of the retina subtending about 5 degrees of visual angle around the center of the viewer's gaze, where the viewer is *fixating*, called the *fovea*. While the characteristics and limitations of foveal vision are very well studied and the design of displays and rendering algorithms reflects this understanding, nearly all the pixels of a high-resolution image fall in the user's peripheral vision at a given moment. The foveal region covers merely 10% of a 15" laptop screen at typical viewing distances, and less than 1% of a modern virtual reality headset. Accounting for loss of acuity in the periphery affords opportunities to reduce computation using *foveated rendering*, (e.g., References [1, 2]). Such systems use gaze tracking and models of certain aspects of peripheral vision to achieve lower rendering or resolution costs while also producing a percept that is identical to that of a full-detail image. As gaze tracking hardware becomes common, foveated techniques have the potential to reduce graphics computation by one or more orders of magnitude—savings that can be channeled into more realistic images, improved power consumption, cost, or form factor for devices, and/or reduced bandwidth across display busses or wireless networks. However, foveated graphics techniques to date leave something on the table: They do not explicitly account for the nature of the *peripheral encoding*, an important source of potential savings. Graphics researchers have largely built on models of peripheral *acuity* that account for the resolution of photoreceptors on the retina [3], as well as for human contrast sensitivity as a function of eccentricity [4]. But, as visualized in Figure 1, human peripheral vision is far from being simply a low-resolution version of foveal vision, and as we discuss below, acuity alone does not account for the limitations and opportunities inherent in foveated rendering.

In recent years, vision science researchers have made great strides toward understanding the peripheral encoding, in particular accounting for the dominant effect in peripheral vision known as *crowding* (see Figure 2 for a classic demonstration). Crowding refers to phenomena in which vision becomes significantly worse when the display or scene is complex or cluttered and occurs even within the central fovea. Although there is a smooth degradation in both acuity and crowding as a function of eccentricity, crowding is the dominant factor in peripheral vision. Recent advances in understanding crowding have been significant, but this new understanding has not yet been operationalized into efficient models that can guide graphics and display tasks such as rendering or compression. Our work advances the state-of-the-art in peripheral encoding by making it faster and easier to develop more sophisticated models that explain performance on a wider range of visual tasks, and for a wider range of visual stimuli, and to test those models in a setting that is more robust, realistic, and relevant to the graphics community. By creating more effective tools for vision science, we hope to pave the way for vision research to better inform graphics pipelines that take full advantage of the limitations of human vision.

Here, we model the peripheral encoding utilized by the human visual system and test this model based on the concept of spatial metamerism. In analogy to metamers studied in color vision, spatial metamers are stimuli that are physically different from each other, but generate the same response in the visual system when viewed at a given eccentricity. One can also talk about, and generate, metamers of a model: *model metamers* are images that are physically different but generate the same response *according to the model*. If a model metamer also serves as a metamer for human vision, then a human subject should be at chance at distinguishing between the synthesized image and the original input to the model. We call this a successful metamer. One can have a successful metamer by simply being conservative about how much one changes the original image, much as one can have a successful foveated rendering by throwing away little information. The goal of a metamer test of a peripheral vision model is to preserve human vision metamerism while throwing away information that peripheral vision does not have access to, much as we would like foveated rendering systems to maximize savings while preserving appearance for the user.

To this end, we present a novel and efficient dataflow model that computes a hypothesized encoding of peripheral imagery to capture significant recent advances from vision science. The model outputs model

Fig. 1. An example of our peripheral encoding model applied to a high-resolution color image for two different gaze locations (outlined in yellow). Under conditions of crowding, observers often describe a "jumbled" appearance, leading to difficulty performing peripheral tasks such as recognizing an object. The output of our model shows jumbling in the periphery, making recognition of details difficult, as often observed in performance of peripheral tasks. Despite their scrambled appearance, at the modeled viewing distance the middle and right images would theoretically be indistinguishable from the original image on the left when focusing on the blue screen and the Hawaiian shirt, respectively. The bottom row shows zoomed-in cutouts of both gaze locations in each image, demonstrating that the crowding model jumbles the peripheral details, rather than decreasing resolution; relatively high resolution details remain, but precise information about their location is lost. We propose an efficient model for generating images that mimic the effect of crowding for use in a variety of graphics and vision science applications. Images are best viewed electronically at full resolution.

G + KG P

Fig. 2. Visual crowding, though applicable to general stimuli, is often demonstrated and studied using letter arrays. Fixating the cross, it is easy to identify the the isolated letter on the left, but hard when that letter is flanked by other nearby letters on the right.

metamers for testing. In addition, we take advantage of our more flexible system to add "end-stopping" mechanisms to the model [5]. Though not a neural network, our model uses the high-performance optimization machinery of modern deep learning to run orders of magnitude faster than the state-of-the-art model. We evaluate our model with a psychophysical study on peripheral texture perception. We believe this work will enable new research on perceptually adaptive graphics, and we hope it also serves as a call to action for further research into and refinement of perceptual encoding models.

Our contributions:

- a survey of current knowledge about peripheral vision's degradation in the face of clutter, including the state-of-the-art model based on measurement of a rich set of summary statistics (Section 2),
- a novel perceptual encoding model that flexibly allows inclusion of new image features such as end stopping, and is geared toward an efficient dataflow implementation (Section 3),
- a new dataset of rendered textures labeled using modernized texture descriptors (Section 5), and
- a psychophysical experiment on texture perception that illustrates the encoding model in action (Section 6).

2 BACKGROUND

As described in the previous section, foveated graphics applications typically use lower-resolution imagery in the periphery to achieve a variety of computational savings. However, vision in the periphery is not simply lower-resolution; the compressive encoding used in peripheral vision is more quirky and interesting. We review

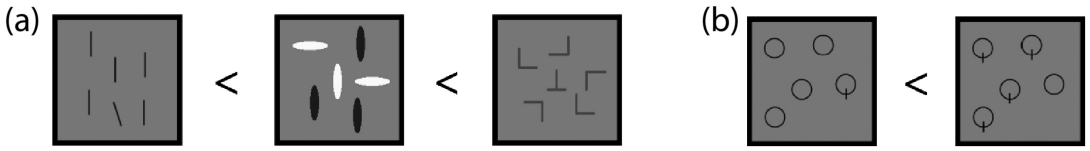

Fig. 3. One of the success stories from modeling crowding has been a new understanding of visual search. Visual search has many peculiarities. (a): Even holding constant the discriminability between the target and other search items, there are big differences in performance, with search for a unique feature (left) usually far faster than search for a conjunction of features (middle, white AND vertical), which is in turn faster than search for a configuration (right, the two bars configured to form a T). (b): A number of asymmetries exist: for instance, it is easier to find a Q among Os than an O among Qs. A pooling model of crowding can explain all of these results [6–8], providing a more parsimonious account than previous theories.

Fig. 4. Peripheral vision lacks detail, largely due to crowding. This lack of detail likely provides some of the explanation for change blindness: As an example, most viewers will have trouble "spotting" the difference between the two images shown here. One of the 128 examples available from Sareen et al. [9].

work from the past few decades exploring a number of odd perceptual phenomena in the periphery, focusing on a simple encoding model (pooling) that explains a surprising amount of the observed behavior. We demonstrate significant computational improvements to (Section 3) and evaluation of (Section 4) this model.

2.1 The Peculiarity of Peripheral Vision

Human vision is not the same everywhere, i.e., across the visual field (see References [10, 11] for reviews). One often hears that peripheral vision has low acuity, or resolution, relative to the fovea, and peripheral vision does more poorly resolve high spatial frequencies. However, it turns out the loss of acuity is actually relatively mild [11]. Furthermore, loss of acuity cannot explain performance on tasks for which the stimuli have typically been designed to avoid acuity limits. Rather, it is the compressive encoding of peripheral vision that drives performance on many visual tasks, even those that do not explicitly probe peripheral mechanisms.

For example, visual search is often difficult [12]: It can be hard to find your keys, even when they are right in front of you. It is easy to tell your keys from your cell phone; if vision were the same everywhere, search should be easy. That search is difficult implies that vision is not the same across the visual field. The visual search literature is full of such phenomena (Figure 3). Even when it is easy to tell the target from other items in the display, some searches are much harder than others, and there are numerous search asymmetries. Human scene perception also demonstrates that vision is not the same across the field of view. A very short glance—in some cases for as little as 30 ms [13]—suffices to get the gist of a scene. However, if we experimentally probe the details of that percept, then we often find that they are murky. One way this can be demonstrated is using a phenomenon known as change blindness, in which observers are asked to find the difference between two similar images, as in a child's puzzle (Figure 4(a)). Viewers also have trouble with two images presented in succession, and this is essentially the same phenomenon as the side-by-side case [14]. The fact that people are often "blind"

Fig. 5. (left) Crowding display from Figure 2. (middle) If loss of peripheral acuity were the cause, then we should be able to blur this image until the crowded "G" on the right were unreadable, but could read the isolated letter on the left. However, the loss of high spatial frequencies affects both approximately equally. Crowding does not merely reflect a lack of peripheral acuity. (right) Applying our metamer model shows the perceptual asymmetry between crowded and uncrowded letters in the periphery.

to the differences between the two images [15] again suggests that vision is not the same everywhere; if it were, then it should be easy to find the differences, as they are easy to see once found.

These phenomena, as well as others, suggest that human visual processing has an information bottleneck. Because of capacity limits, one strategy that the visual system uses is to more coarsely encode information farther from the point of fixation. This strategy degrades information where you are not looking—i.e., at any given moment, almost everywhere. The nature of this encoding and the information that survives the bottleneck amounts to degradation in the face of clutter, known as crowding. Crowding has often been demonstrated using arrays of letters like that shown in Figure 5. Fixating the central "+" at a normal reading distance, one likely has no difficulty reading the isolated "G" on the left. Add additional clutter, the flankers "K" and "P," and it becomes much harder to read the crowded "G" on the right. Move those flankers farther from the target letter, and at some "critical spacing" the task becomes easy again [16]. We know that crowding is not due to the loss of peripheral acuity [17]. One can easily demonstrate this (Figure 5, middle). Fixating the crowded "G" on the right, we can progressively blur the image until it becomes difficult to read the letter. However, at that level of blur it also becomes hard to read the uncrowded "G" on the left. Furthermore, examining the blurred image it is obvious that even at this excessively low resolution the flankers do not interfere with perception of the central target. Rather, there must be some other kind of loss of information in peripheral vision. Furthermore, this loss does not apply merely to letter arrays, but rather to perception of any visual stimuli [11].

The phenomenology of crowding has given us some hints as to the underlying mechanisms. When viewing a cluttered or complex peripheral stimulus, firm localization of detail becomes difficult [18] (as translated by Reference [19]) [20–22]. Lettvin [23] remarked that "It is not as if these things go out of focus—but rather it's as if somehow they lose the quality of 'form.'" This loss of location and form information can lead to a percept one might describe as "jumbled." A peripherally viewed word "only seems to have a 'statistical' existence... [preserving] every property save that of the spatial order that would confer shape." Peripheral vision tolerates considerable image variation without giving us much sense that something is wrong, despite such variation appearing blatantly obvious when viewed foveally [24, 25]. A number of recent review papers give more detailed description of these and other crowding phenomena [26–28].

2.2 The Summary Statistic Encoding Hypothesis

These phenomena—the jumbled percept, loss of location information, and the seemingly statistical nature of the perceived stimulus—have pointed a number of researchers toward a particular explanation. Crowding has been attributed to "excessive or faulty feature integration," "compulsory averaging," "forced texture processing," or "a statistical representation" [23, 26, 27, 29, 30]. This led to the suggestion that crowding results from an encoding scheme that *pools summary statistics over local regions* [30, 31], and to the specification and implementation of an image-computable model [24, 29, 32].

One can make a number of arguments for why the visual system might implement such an encoding. Averaging in the sense of collecting a rich set of *summary statistics* preserves a great deal of useful information, at the cost of discarding the precise phase and location of details [29]. Take, for instance, the 700–1,000 summary

Fig. 6. Visualization of a partial lattice of pooling regions overlaid on an image, with a viewing position centered on the middle penguin. As one moves from the center of fixation into peripheral vision, pooling regions become progressively larger, integrating information over larger regions; at the edges of the image multiple objects fit within a given pooling region. For visualization clarity, the pooling regions shown are sparser than those used to create our metamers, and the central modeled foveal region has been shown without pooling.

statistics used by Portilla and Simoncelli's [33] state-of-the-art model of texture appearance (note the exact number depends on the choice of several model parameters). Those statistics include such things as the distribution of luminance and color, as well as pairwise cross-correlations applied to the output of oriented filters at multiple scales, similar to mechanisms found in early visual processing areas. These latter statistics to some degree encode, for instance, the presence of extended contours, periodic structures, corners or other junctions, as well as the sharpness of edges. Portilla and Simoncelli [33] used their texture analysis/synthesis procedure to generate new samples of "texture" that have approximately the same summary statistics as the original, demonstrating that these summary statistics do a good job of capturing the appearance of textures. This type of statistical textural representation provides an efficient encoding, of use for getting around an information bottleneck in vision [29].

Furthermore, considerable work in vision science has suggested that many visual tasks are inherently statistical tasks. Texture segmentation and discrimination rely on summary statistics [34]. Deviation from local summary statistics makes an unusual item *pop out*, seeming to draw the observer's attention [35–37]. Deciding whether a material is shiny might have to do with subband skew [38]. Real-world scenes contain many textured regions—trees, sky, building façades, stone walkways—and representing those textures well may facilitate scene perception [39]. A statistical encoding in peripheral vision, then, may support many real-world tasks.

2.3 Image Synthesis from Summary Statistics

To model peripheral crowding, according to this hypothesis, one needs to capture the loss of information in peripheral vision due to computation of a rich set of summary statistics [24, 32]. These summary statistics are computed within each of multiple pooling regions that grow linearly in size with *eccentricity* (i.e., distance to the point of fixation), overlap, and densely tile the visual field (Figures 6 and 7). The state-of-the-art model of peripheral crowding uses approximately the summary statistics of Portilla and Simoncelli's [33] model of human texture perception. This model has largely been tested by asking to what degree the hypothesized encoding can predict performance at visual tasks, rather than by testing its ability to generate metamers for the visual system. The loss of information in the hypothesized peripheral encoding can predict difficulty recognizing peripheral objects in cluttered displays or scenes [6, 7, 24, 29, 40]. The loss of information also predicts difficult search conditions, while preserving the information necessary to predict easy search—conditions in which the target appears to "pop out" and draw attention [6–8]. In spite of this loss of information, the encoding preserves sufficient information to predict the ease with which observers get the gist of a scene at a glance, including identifying the scene category, upcoming turns when driving, and the presence of target objects like an animal or a stop sign [41, 42].

Questions remain regarding the correct set of summary statistics to use. The state-of-the-art model, following Portilla and Simoncelli [33], uses hand-picked statistics chosen based on understanding of the utility of their

Fig. 7. A metamer of the image used in Figure 6 generated with our peripheral encoding model using the same fixation point. When centering gaze at this fixation, the summary statistics under an ideal model should be the same as in the original within each pooling region, and distortions at image edges would be undetectable in peripheral view. This is despite the extremely distorted appearance of the image away from the modeled fixation when viewing those regions foveally.

features, efficiency at encoding natural images, and ability to capture texture appearance. Little work has been done to test alternative statistics for modeling peripheral vision. Given the recent success of deep learning, one tempting possibility consists of learning the statistics by training a network to perform a visual task. Researchers have begun to explore this possibility. Gatys et al. [43] and Wallis et al. [44] use the intermediate activations of a pre-trained CNN (VGG-19) pooled over the entire image. This amounts to roughly 10 to 100 times more statistics. Compared to Portilla and Simoncelli [33], the resulting textures better mimic the originals when viewed foveally, but achieve roughly equal quality when viewed peripherally [44]. While it is promising to derive summary statistics from a CNN trained to perform core object recognition, rather than hardcoding them as in Reference [33], considerably more work needs to be done to test the usefulness of those statistics in modeling peripheral vision.

Additionally, somewhat different methods exist that mimic aspects of peripheral information loss. For example, Deza et al. [45] propose a learned manifold of image features relative to a universal metamer noise image. In effect, they have learned the ways in which they can distort an image while maintaining a degree of metamerism, with the amount of distortion able to vary as a function of eccentricity. Walton et al. [46] also propose a fast way to render visual metamers for graphics applications. Their method uses very simple statistics: Per-pixel mean and variance are computed on the output of a steerable pyramid with a single filter size and two orientations. This approach allows for faster (real-time) computing of metamers. However, the simplicity of their model is unlikely to correctly account for the complex crowding phenomena described earlier in this section; one way to ensure metamerism is to throw away less peripheral information. As with previous work, they were also only able to validate their method using a small number of natural images (seven), while our work examines 200 different textures to determine which natural image consituents might be well-represented by our model.

As in the demonstration in Figure 5, one can gain intuition about what information is preserved and lost in peripheral vision by synthesizing images with approximately the same summary statistics. Much of the previous work [24, 32, 44] has done this by building upon texture synthesis techniques, e.g., References [33, 43]. Portilla and Simoncelli [33], for instance, pool summary statistics over a single pooling region that covers the entire input image. They then synthesize a new texture by starting with a random image and iteratively applying the summary statistics as constraints, adjusting the synthesized image until it converges to have approximately the same statistics as those measured in the original.

Texture synthesis techniques can be extended to model the progressive loss of information in peripheral vision by satisfying statistical constraints computed not over the entire input image, but rather computed within multiple, overlapping pooling regions (such as the pooling regions visualized in Figure 6). This process synthesizes model metamers for a given model of peripheral vision. The success of such model metamers can be tested by assessing a human subject's ability to distinguish it from the original input image; successful metamers are indistinguishable from the original image.

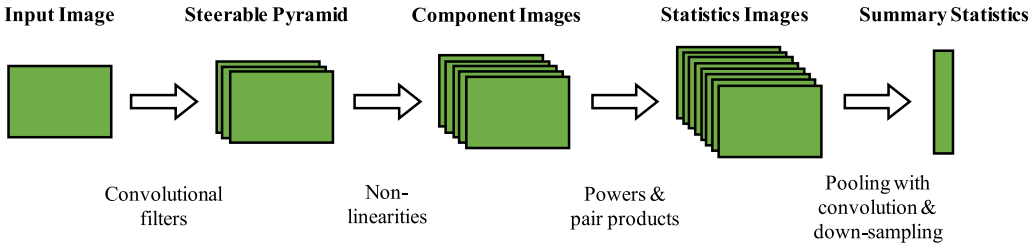

Fig. 8. Dataflow for computing summary statistics for an image. First, we build a steerable pyramid using linear convolutional filters to get an initial set of component images, corresponding to different scales and orientations. These include low-pass, high-pass, and edge phase (or oriented band-pass) images. Additional component images (edge magnitude and edge phase-doubled) are created using non-linear operations to create an augmented steerable pyramid. Then, we create statistic images that correspond to various moments and correlations of the component images. Generally these are computed using pixel-wise operations, exponentiation, or multiplying two component images together, followed by pooling by applying a convolutional blurring filter to the statistics images and downsampling them. The results contain the statistic values for each pooling region, which we concatenate into a list. To generate a model metamer, we first compute the statistics list for a target image and then use a gradient-based optimizer to solve for another image with matching summary statistics.

Directly extending Portilla and Simoncelli [33] to multiple regions using a related Fourier approach [24] is quite computationally intensive, however, requiring on the order of 6 hours to synthesize one image from a 512×512 original. We use a different approach, aimed at efficiency and flexibility, utilizing a dataflow network with the desired statistics wired into it (rather than being learned or computed in a Fourier space), as detailed in the next section.

3 PERCEPTUAL ENCODING MODEL

In the peripheral encoding model an image is represented by a set of image statistics that have been averaged over local pooling regions. If two images have matching pooled statistics, then the model predicts they will be visual metamers. Loosely speaking, filtering with a multiscale, multiorientation pyramid encodes the presence of local "features," while the pooling removes information about the location of these features. Thus, two model metamers will contain similar features but potentially rearranged or jumbled within the pooling regions. This approach to modeling peripheral crowding was pioneered in the work of References [24, 29, 41].

In our implementation, the pooled statistics are computed using three stages. First the input is converted into a set of component images split by color channel, spatial scale, and orientation. The component images are based on the subbands from Steerable Pyramids [47], augmented by non-linear operations (computing statistics as in Reference [33]). Next, we compute statistic images by taking paired products and powers of the component images, essentially computing a number of moments and correlations. Then each statistic is locally averaged to compute a single value for each pooling region, effectively blurring and downsampling the statistic images. As with prior methods, our statistics are based on the texture synthesis work of Portilla and Simoncelli [33], though we use a modified set of statistics, as we will discuss later. The result is stored as a stack of low-resolution pooled images, one for each summary statistic. Figure 8 illustrates the dataflow for this process. Typically, we use about 300 statistics for grayscale images and roughly 1,000 for color images. For comparison, Freeman and Simoncelli [24] and Rosenholtz et al. [41] use between 700 and 1,000 statistics for grayscale images. A more detailed breakdown of statistics by category is included in the Appendix. Note that this model only simulates the effects of pooling and does not model loss of acuity as a function of eccentricity. One can easily add the loss of acuity as front-end eccentricity-dependent filtering applied to the input image [48].

To create model metamer images, we first compute the pooled statistics for a target image. Then, we iteratively adjust a seed image using a gradient-based optimization process until its pooled statistics closely match those of

Original Image	F&S statistics Matlab code	F&S statistics Our system	Our statistics Our system
Time (minutes)	188 m	6.6 m	1.7 m

Fig. 9. Sample results generated using the Matlab code provided by Freeman and Simoncelli and from our new system. The first image is an example they provide. The other three are randomly generated model metamers with the gaze point at center. The second image is generated by the Freeman and Simoncelli code using its default settings and took 188 minutes. The third image is generated by our system, configured to match their pooling and statistics as closely as we could, and took 6.6 minutes. The two systems handle pooling regions that extend beyond the image differently, leading to larger differences there, but are otherwise of similar quality. The third image is generated with the same pooling regions but using our reduced set of statistics and took 1.7 minutes to compute. Using our statistics is both faster and produces a higher-quality result. Images generated using a 3.5 GHz Intel i9-9900X CPU with an Nvidia RTX 2080 Ti GPU. The Matlab code uses the CPU only while our system also uses the GPU.

the target image. Our system is implemented using the PyTorch library [49], which provides automatic gradient computation, highly optimized image operations, and the ability to easily utilize GPUs. These features have allowed us to work with higher-resolution images and greater freedom to experiment with different statistics as compared to prior work. Using automatic differentiation allows us to easily modify our statistics without having to invent new optimization routines for each type of statistic, as in Reference [33]. Our system's structure resembles a convolutional neural network, though in our case, the network is fixed and the input pixels are learned through gradient descent to match desired output statistics (input image is learned, weights are fixed). A comparison of the output from our system and Freeman and Simoncelli's reference Matlab implementation is shown in Figure 9. We plan to release our system as open source.

3.1 Component Images

For color images, we first split the image into color channels. We use a perceptually based three-channel opponent color space, which consists of an achromatic, a red-green, and a blue-yellow channel [50]. For grayscale images only the achromatic channel is used. Each channel is then handled mostly independently and identically in our system, though we do include a few cross-channel color correlation statistics, as described in the next subsection. There is relatively little discussion of the role of color in the prior work. Freeman and Simoncelli's [24] sample code uses a custom per-image color space based on a **principal components analysis (PCA)**, while Rosenholtz [32] used **independent components analysis (ICA)**. These per-image color spaces are content-dependent and harder to extend beyond still images. We found that using opponent color channels is simpler and works well across a variety of images.

Next, for each channel, we construct an augmented steerable pyramid in the same style as in Portilla and Simoncelli [33]. A steerable pyramid uses a set of convolutional filters tuned to different spatial scales and orientations to split the image into corresponding components. Most of these filters resemble Gabor-filters and are very good at detecting edge-like features in the images. Steerable pyramids can be viewed as a soft partitioning of the image in Fourier space and can be computed using Fourier transforms for efficiency. The augmented

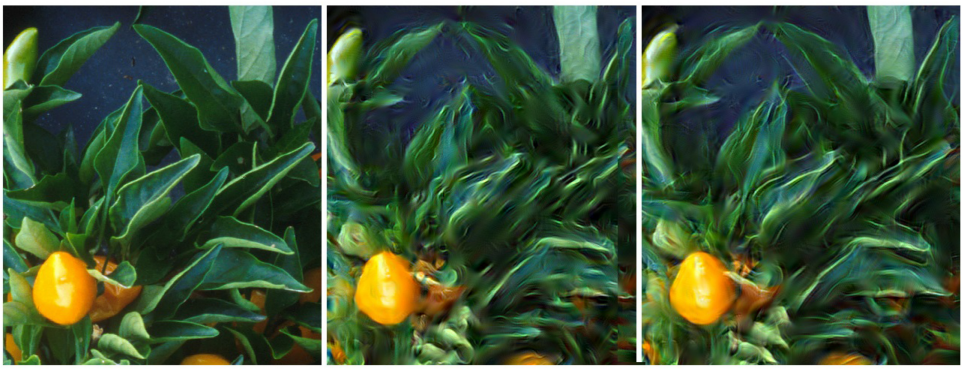

Fig. 10. Including two additional filter orientations and end-stopping statistics noticeably improves the quality of the generated model metamers. Shown are insets from the original (left), 4-orientation/non-end-stopped model metamer (middle), and model metamer generated with 6-orientations and end-stopped statistics (right). These two additions greatly improve the ability of the model to reproduce high-contrast, continuous, curved edges. Note that while these statistics improve the leafy section of the image, they have little effect on the yellow pepper in the bottom left. Original image from Reference [51].

steerable pyramids contain a few extra components. Each edge filter is replaced with a quadrature pair (filters offset by a 90-degree phase shift) [47] to better track edge polarity and correlations. The quadrature pairs are also combined to create edge magnitude images. In addition, Portilla and Simoncelli [33] added phase-doubled images for tracking cross-scale correlations. We use the same components as Portilla and Simoncelli [33], except that we extended the pyramids to use 6 spatial levels instead of 4, and use 6 orientations instead of 4 (Figure 10). These changes allow us to create higher-quality model metamers for images with features at a wider range of scales and for images with strongly directional patterns.

3.2 Image Statistics

The image statistics correspond to various moments and correlations of the component images. (The final step in computation of these summary statistics involves pooling over the pooling regions described in the next subsection.) Our statistics are based on those of Portilla and Simoncelli [33], with some significant changes. For computational simplicity, we use raw rather than central statistics (i.e., we omit the mean subtraction), as the mean values are also included among the statistics to be matched. We also removed some statistic types that did not seem to have much effect on metamer quality in our preliminary tests. In particular, we removed the low-pass component statistics and all of the autocorrelation statistics. This is a considerable savings, as numerically these constituted a majority of the original statistics. Portilla and Simoncelli [33] originally included these statistics as necessary to synthesize a texture from summary statistics pooled over the entire image, but with multiple pooling regions these statistics may be redundant. The remaining statistics for each color channel image include: pixel mean value, edge magnitude mean and variance at each scale and orientation, correlations of edge magnitude across orientations at the same scale, and correlations of edge magnitude and phase between neighboring scales at the same orientation.

For color images, we add a few cross-color-channel correlations, namely, between edge magnitudes and between phase components at the same scale and orientation. These are important for reproducing multi-channel hues such as orange or purple. The use of opponent color channels along with these few extra statistics has worked quite well in our tests across a wide range of input color images.

Nearly all our statistics are edge-based and respond most strongly to linear edge features. This corresponds roughly to how simple and complex neurons work in human visual cortex [52]. The early visual system also

has neurons with other properties, such as end-stopped neurons (originally called hypercomplex cells [53]), that respond most strongly only where edges end, curve, or change direction. End-stopping has been shown to be important in image perception [54, 55] and argued to be important for efficient encoding [56], but is not well represented in the prior statistics. Thus, we have added new statistics that better capture end-stopping by measuring the change in strength along an edge. These are computed by subtracting each edge magnitude component image from a copy of itself translated a short distance along the expected edge direction (i.e., the filter orientation) and then squaring the result. In our tests, these new end-stop statistics improved metamerism in regions with lines that curve or end (Figure 10).

3.3 Pooling Regions

Each image is conceptually subdivided into a set of overlapping pooling regions. Within each pooling region only the average value of each statistic is kept. This is intended to model some effects of the pooling of neural inputs that occurs during processing in human visual cortex. As discussed in Section 2, summary statistic models hypothesize that many limitations in our peripheral vision, such as crowding, can be explained by such pooling. Consistent with Bouma's Law for the critical spacing in crowding [16, 27, 57], the size of the neural pooling is believed to increase roughly linearly with angular distance from the gaze point. Thus, the maximum acceptable pooling region size in our model depends on the eccentricity, i.e., the visual angle from the point of gaze. However, this is only an upper bound on the acceptable computational pooling sizes; using smaller regions, which enforces closer match to the input image, should also produce visual metamers.

Our system can generate two different patterns for the pooling regions: uniform and gaze-centric. Uniform model metamers use the same size pooling region everywhere but nonetheless may be successful metamers if viewed beyond the eccentricity corresponding to that pooling region size. In gaze-centric model metamers, the pooling size varies linearly with eccentricity and thus relative to a specific gaze point. For efficiency, we compute gaze-centric model metamers by first warping the image into a log-polar space to equalize the pooling region sizes and shapes, then computing a uniform model metamer in this space, and finally transforming the result back to normal image space. The axes of the log-polar space are the logarithm of the radius from the gaze point and azimuthal angle around the gaze point with scaling factors chosen so the transformed pooling regions are equal-sized circular regions of a chosen size.

In our system the pooling involves convolving the statistic images with a pooling kernel, effectively blurring them, and then downsampling the result to reflect the limited number of pooling regions. The results are low-resolution pooled statistic images where each pixel represents one pooled statistic in one pooling region. Our pooling kernels are based on Freeman and Simoncelli [24] with a few significant changes. We use pooling regions that are approximately circular in log-polar space rather than elongated ones, though sized to have approximately the same area as in their V2 model. We also use a higher density of pooling regions with much more overlap between them. The spacing between neighboring pooling regions in their system was 3/4 of their diameters but only 1/4 of the diameter in ours. Higher density means the model preserves more information about the original image. However, our density is likely higher than actually needed, and we plan to try reducing it in future work.

The supplemental material includes a video showing application of our model to dynamic viewing conditions and a short dynamic movie with static gaze. The video also shows an example of the convergence of our iterative solver.

4 VALIDATION OF THE MODEL

Our new dataflow model enables experiments that could not easily have been run before. Many experiments require generating a large number of model metamers. For example, if one wanted to test whether a peripheral encoding model could predict performance on scene perception tasks, or if one wanted to test the degree to which model syntheses were indistinguishable from the original images, one would need to synthesize a number

of model metamers for each original in a large set of scenes. At absolute minimium, this would require synthesizing hundreds of images. Synthesizing 400 grayscale images at 640 × 480 resolution would take as much as 6 *hrs* × 400 = 100 days. This is a big commitment to run a simple experiment! As a result, most previous work has used a very small number of images. For example, Ehinger and Rosenholtz [42] used a cluster of computers to synthesize 400 grayscale images in a "mere" week, to test whether the information in peripheral vision could explain difficulty performing scene perception tasks. Freeman and Simoncelli [24] tested scene metamerism (whether you could tell one metamer from another while fixating) with a mere 4 images. Wallis et al. [58] tested scene metamerism with a mere 20 images. Alternatively, some researchers have used a CNN-based model of similar architecture to the summary statistic model to synthesize large numbers of model metamers (e.g., Reference [58]). This has the advantage of computational efficiency, but the disadvantage that, due to dramatic differences in the early stages of the models, we do not know whether such CNN-based models actually model peripheral vision.

With our new dataflow model, we can synthesize 400 full-color 1,920 × 1,080 images on a single GPU in a more manageable 23 hours, facilitating many experiments. Here, we use this capability to study metamerism in texture perception, which is relevant in many graphics applications. For example, video games often use variable level of detail for both scene geometry and object textures (i.e., mip-mapping [59]) in distant regions away from likely fixation, and some foveated rendering techniques such as fCPS [60] explicitly use texture downsampling in the periphery. Even within a complex scene, certain objects (i.e., trees) can be thought of as textures in terms of the image that is projected onto the eye [39], indicating that our findings should have some application for metamerized scenes as well.

4.1 Discrimination of Synthetic vs. Real Textures

Previous work has examined whether or under what conditions participants can distinguish between original and synthesized textures shown in the periphery. Much of this work was restricted to texture perception in the near-periphery, or collected statistics in a non-foveated way, over the entire texture patch, using the Portilla and Simoncelli [33] texture synthesis model. As a result, this work tested not true metamerism, but rather, essentially, whether the original and synthesized textures appear sampled from the same kind of texture. The advantage of using a non-foveated model is speed (about 2 minutes to synthesize a 256 × 256 patch). However, collecting statistics over the entire patch makes no distinction between foveal and peripheral perception, discarding the same information regardless of eccentricity of the patch, and does not make use of the information available to peripheral vision from neighboring, overlapping patches. This disadvantages non-homogeneous textures and those with larger scale structures and makes it difficult to interpret the results in terms of the encoding in peripheral vision. Using a non-foveated model also disadvantages textures shown near fixation. Balas [61] examined 15 textures from the Brodatz database [62] and asked which statistics measured in Portilla and Simoncelli [33] were necessary and sufficient for capturing appearance for those textures. Textures as displayed to the participants extended from near the fovea out into the periphery; due to the statistics being pooled over the entire image, artifacts in or near the fovea may have allowed participants to distinguish between original and synthesized textures. Keshvari and Wijntjes [63] studied material identification in (1) a 2 deg diameter patch of original texture at fixation; (2) a 2 deg diameter metamer synthesized using Portilla and Simoncelli statistics measured over the image as a whole, also shown at fixation; or (3) a 2 deg diameter patch of original texture at 10 deg eccentricity. Their test set included 50 images in each of 6 material categories: fabric, foliage, stone, water, and wood. They found that material perception with the synthesized images predicted peripheral but not foveal material perception with the original images.

Wallis et al. [58] examined the foveated encoding of Freeman and Simoncelli [24] (pooling of statistics over multiple overlapping pooling regions) for 10 "texture-like" images and found that observers had difficulty telling apart original textures from synthetic. However, observers could reliably distinguish more "scene-like" images from Freeman and Simoncelli [24] metamers and concluded that a pooling model is insufficient, proposing adding

content-dependent grouping and segmentation. However, it could also be that the Freeman and Simoncelli [24] model metamers would be successful metamers if they encoded more summary statistics, or different ones, or that the implementation has difficulty converging to the point at which the summary statistics capture grouping information.

4.2 Our Approach

Our experimental design differs significantly from that used in most previous work. First, we leverage the speed of our system to synthesize and present high resolution color stimuli with a larger field of view than any previous study on summary statistic metamers (29 deg vs. at most 25 deg [58], and often much less). Additionally, we increased the stimulus duration from 300 ms to 2.5 sec and presented stimuli side-by-side rather than sequentially in time. These changes require that the model be correct at a much broader range of eccentricities and for a longer period of sustained peripheral attention, setting a more stringent (and arguably more realistic) criterion for a successful metamer. Simultaneous presentation of both model metamer and original ensures that the texture boundary between the two is also unnoticeable and provides a better setup for performing the same experiment with video.

Additionally, to study participants' ability to distinguish metamer textures from the original, we curated our own dataset of 400 texture images (described below). All summary statistic models (including ours) create better visual metamers for some images than for others. Examining a wide variety of textures allows us to begin to assess what image characteristics determine whether a metamer succeeds, i.e., is indistinguishable from the original given a specific viewing distance and direction. Differences in metamerism did not seem to correlate with the residual loss of the optimization procedure, suggesting that at least some statistical properties of the images that participants can perceive in peripheral vision were not captured by the model. Our goal therefore was to test many different textures to determine which textural qualities might not be well-represented by our statistical descriptors. We also analyze our results in terms of material category to gain additional insight.

5 RENDERED TEXTURE DATASET

We curated a dataset of 190 **physically based rendering (PBR)** textures that cover a range of material categories: fabric, fire, foliage, food, fur, glass-like, ground-like, leather, marbled, metal, painted, paper, plastic, rock, snow- or smoke-like, tiled, water, wood, and an additional "other" category for interesting textures that did not fit into any of the material groups above. The textures were hand-curated across 15 online databases of high-quality free-use PBR materials with the goal of covering both the most commonly used material categories and a representative sample of scale and shape within each category. We sought both to span the range of natural textures and to include texture materials and qualities relevant to graphic artists. The material categories chosen (Figure 11) represent the union of previous texture material databases [64] and online texture databases that characterize materials by other qualitative features [65–79].

Each texture was rendered using G3D [80] at half native resolution tiled onto a $1{,}920 \times 1{,}080$ RGB output image. Textures were rendered from a fronto-parallel angle under two different cube map lighting environments (white room and plain sky) coupled with two different angles of directional lighting (oblique and direct, respectively). We also included 20 high-resolution photographs of clouds (from Reference [81] and the authors' personal collections), since clouds are commonly used as background textures in many rendered outdoor scenes. The total size of the dataset was therefore 400 images, 20 images per category across 20 categories.

5.1 Categorical Texture Descriptors

The size and diversity of our texture dataset represents a broad swath of realistic textures, but analyzing the effectiveness of our model for each texture individually would be unwieldy and prevent generalization to new textures. We therefore sought to find unifying textural descriptors that would provide graphics users of our

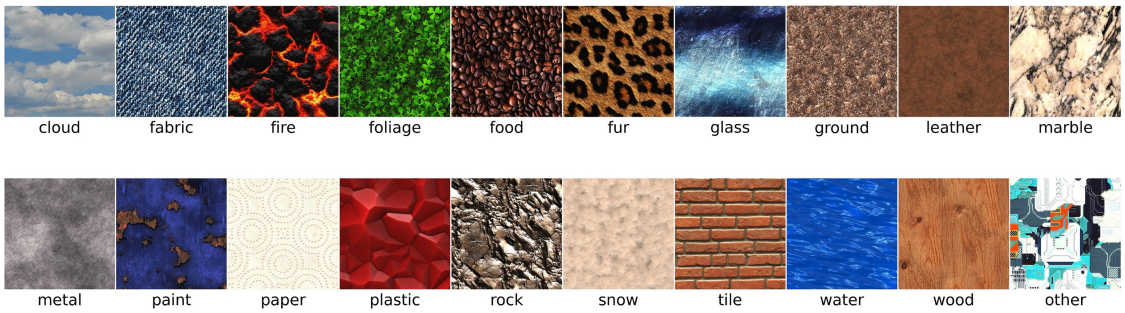

Fig. 11. An exemplar texture for each of the 20 material categories in our dataset. Our set of 400 textures spans a range of material descriptors. Contrast enhanced for better viewing at a small scale.

peripheral encoding model an intuitive understanding of which textures make good metamers and which do not. There have been many efforts to develop automated texture classification methods [82], but the academic community lacks consensus on which intuitive categorical texture descriptors are the right ones to use. Based on multidimensional scaling of human judgments, Rao and Lohse [83] proposed three dimensions: repetitive vs. non-repetitive; high contrast non-directional vs. low contrast directional; and coarse low complexity vs. non-granular, fine, and high complexity. However, Liu and Picard [84] suggested a different, simpler set of three texture descriptors: periodicity, directionality, and randomness. Another set, of six descriptors, was proposed by Tamura et al. [85] and is still widely used to characterize images and textures [86]. Their proposed texture characteristics are coarseness, contrast, directionality, line-likeness, roughness, and regularity.

We opted to follow the general direction of Tamura et al.'s [85] texture descriptors, but updated them to reflect modern practices in both graphics and vision science, and to fix issues they themselves identified. In this article, we propose a new set of computational measures that we believe match the intuitive definitions of the original six descriptors from their 1978 paper. The Appendix includes a full description of our implementation for each descriptor, as well as examples of textures corresponding to the maximum, median, and minimum values for each descriptor. We offer these updated texture descriptors as a contribution to the community, and the code is available at https://github.com/vdutell/TextureFeaturesPaper.

6 PSYCHOPHYSICAL EVALUATION

Due to restrictions on in-person data collection, we conducted our experiment online using the Psychopy [87] and PsychoJS libraries on the Pavlovia server platform. The Appendix includes some caveats related to remote studies and our efforts to mitigate them. Thirty-two adults participated in the experiment, although 3 were excluded based on an excessive number of timeouts and late responses (details below). The remaining 29 participants ranged in age from 24 to 54 (mean of 35), and 9 reported wearing corrective lenses. Participants were asked to use a desktop monitor, not a laptop screen, in a dimly lit room, to maximize display visibility and contrast. Recruitment of subjects and administration of the study was conducted in accordance with the Declaration of Helsinki Ethical Principles for Medical Research Involving Human Subjects.

On each trial, participants were shown a full-screen image of a texture at 1,920 × 1,080 resolution, with the screen divided into three sections and a fixation cross at the center, as depicted in Figure 12. Either the top-left or top-right side of the screen showed a model metamer of the texture, while the other side showed the original image. The order of the sides was randomized across textures, such that half of all trials showed the model metamer on the left and vice versa. The lower section of the image showed a model metamer on half of the trials and the original for the other half, randomly chosen. Subjects were instructed to maintain fixation at center and respond as fast as possible whether the bottom section matched the left or right side of the screen. The matching

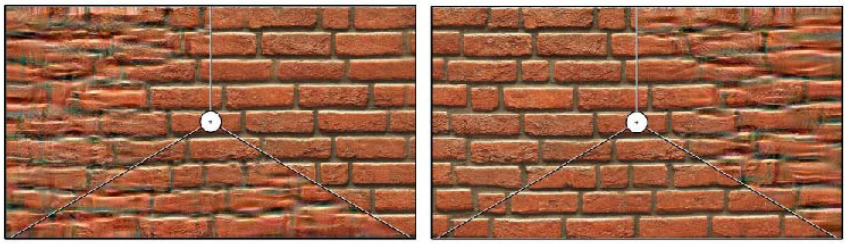

Fig. 12. Two example stimuli from the experiment. Each of these images would have been shown full screen. Participants were asked to fixate on the "+" in the center and indicate whether the lower section matched the left or right. For both examples shown here, the correct answer is "LEFT."

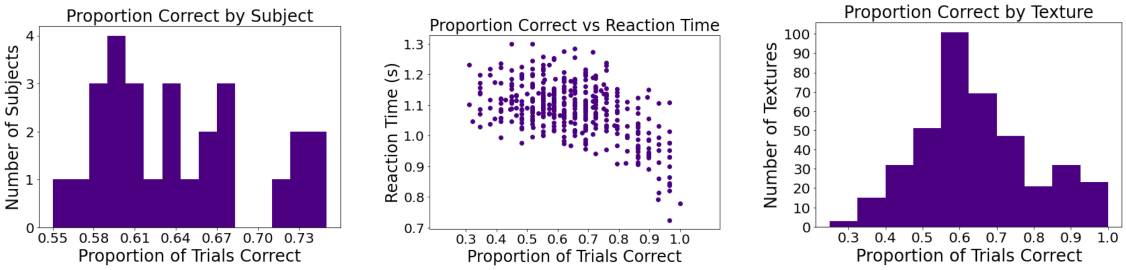

Fig. 13. (left) Histogram of proportion correct for each subject across entire texture dataset. Note that all subjects are above chance, while none reach greater than 75% correct overall; (center) proportion correct vs. reaction time. Each data point represents a single texture/model metamer pair; and (right) histogram of proportion correct for each texture, across all subjects. Horizontal axes for all plots are scaled to fit the data.

sections could both be original texture images or both model metamers. Participants used the left and right arrow keys to indicate their response. We recorded both accuracy and reaction time.

Participants were instructed on the experimental procedure at the start of the experiment, including a demonstration of the feedback sounds and eight practice trials. All practice trials had to be answered correctly before moving on to the main experiment. Trials proceeded in blocks of 10, with rest breaks after each block to remind participants of the instructions and encourage better fixation. If participants failed to respond within 2.5 seconds, then they would hear a feedback sound indicating they had timed out, and the trial would be aborted and repeated at a random point later in the experiment. Furthermore, in the small percentage of the trials in which participants timed out, they could have responded late and inadvertently answered the next trial incorrectly. We therefore discarded any responses within the first 200 ms of each trial.

6.1 Results

Figure 13 shows a summary of our results by subject and by texture. Across participants (left), we observed a wide range of skill levels in identifying synthesized images as metamers, from 53% to 79% with an average of 64%. We also saw a range of average reaction times across subjects, from 0.63 sec to 1.57 sec, with a mean of 1.08 sec. In general, reaction time showed a moderate Spearman's Rank correlation with percent correct ($\rho = -0.416$, $p \ll 0.01$), where longer reaction times correlated with lower percentage correct, as expected. Figure 13 (center) shows the relationship between reaction time and proportion correct. As mentioned above, three participants were excluded for having too many timeouts (number of timeouts greater than two standard deviations from

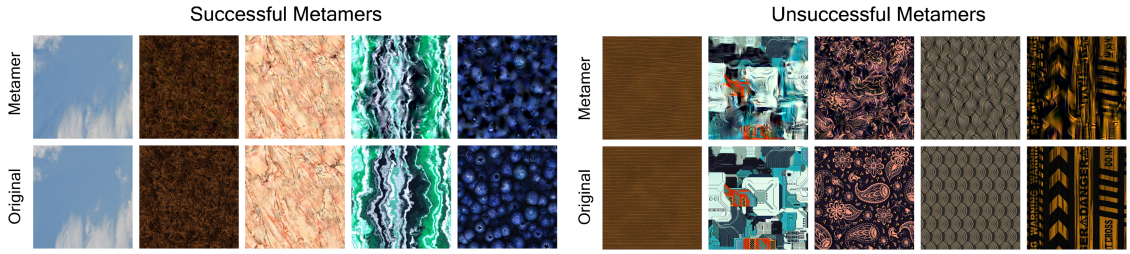

Fig. 14. (left) Selected examples of original and synthesized textures that succeeded, i.e., with participants at chance for discriminating original from model metamers, and (right) examples of textures that failed, with participants above chance at distinguishing synthesis from original. Some of the successful metamers, such as the clouds and pink granite, appear as near-metamers even foveally, while others such as the blueberries are metamers only in the periphery. Shown are 512×512 sub-images taken at half maximum eccentricity. Best viewed electronically at full resolution.

the mean). For the remaining participants the average number of timeouts was 5.5, and the average number of accidental late responses was 2.3 (defined as a reaction time less than 200 ms). Consistent with our expectations, the distribution of mean accuracy per texture (Figure 13, right) also covered a wide range, from 31% to 100% correct, with an average of 64% correct. Figure 14 shows examples of "successful metamers," or textures that were difficult to distinguish from their model metamers (accuracy indistinguishable from chance level of 50%) and also "unsuccessful metamer" textures that were easy to distinguish (average accuracy above 90%).

Note that performing this experiment online without fixation monitoring likely underestimates how many syntheses are successful metamers. While poor lighting or monitor calibration might lead to poorer discrimination between original and synthesized textures, other issues with online experiments—most notably, lack of eye tracking to validate a subject's fixation during each trial (see Appendix)—would only result in more trials in which participants can distinguish between original and synthesized images, and therefore a lower number of textures that succeed as metamers.

We were initially surprised at the below-chance performance for some model metamers (less than 40% correct across subjects). Below-chance performance might suggest that the participant can distinguish between the original and model metamer; does this imply that these model metamers poorly capture the appearance of the original texture, but for some reason participants performed the task incorrectly? We examined model metamer-original pairs leading to below-chance performance and found it extremely difficult to discriminate between the two even when fixating (except perhaps in regions modeled as more eccentric, where artifacts are more visible both theoretically and in practice). However, many of these examples had inhomogeneities in the texture that may have caused observers to consistently and incorrectly pair original with model metamer. Faced with a model metamer difficult to discriminate from the original, an observer may have found, for instance, a patch in the original with an unusual orientation and looked for that orientation in the other regions. If by chance one of the model metamer regions also contained that unusual orientation, then the observer might incorrectly match model metamer with original based on this accidental similarity between the two. We suggest, then, that the below-chance model metamers are successful metamers, and in fact are so difficult to distinguish from the original that observers relied on other, irrelevant differences arising from inhomogeneities in the textures.

6.2 Analysis

Our image dataset spans a range of texture properties, and as anticipated, these results establish that our model encodes some of these properties well and some less well (see Appendix). To begin to get a sense of what improvements one might make to the model, we can next look for commonalities among the textures leading to successful metamers (those that our model replicated sufficiently well to mimic the original in the periphery),

as well as among the textures leading to unsuccessful metamers (those for which our model did not successfully mimic the original). This analysis could also motivate further psychophysical experiments on texture perception in peripheral vision to further explore relevant factors.

We first asked whether material category alone predicts which textures were easier than others to distinguish from their model metamers. Material categories differed greatly in mean accuracy. On average, the most difficult materials to discriminate from their synthesized version were snow/smoke (49% correct) and painted textures (53% correct), and the easiest materials to detect were plastic (81% correct) and tiled textures (87% correct). To determine whether the differences across material categories were significantly different from chance, we calculated the average distribution of accuracy across 100 sets of 20 randomly selected textures and compared this random average to the distribution of accuracy per category using a Mann-Whitney Rank test [88]. None of the material categories were significantly different from chance (lowest $p = 0.03$, with a Bonferroni-corrected significance level of $p < 0.0025$), likely due to the deliberately large variation of appearance within each category.

Additionally, we examined the relationship between accuracy and lighting. Our expectation was that side-lit textures would on-average have higher contrast and different edge and spatial frequency distributions due to self-shadowing, and we wanted to ensure that our texture representation did not confer an advantage for one lighting condition over another. We found a moderately high Spearman's Rank correlation between accuracy across the two lighting conditions ($\rho = 0.623$, $p \ll 0.001$); the average percent correct for each lighting condition was 64.7% (front-lit) and 64.6% (side-lit). A Mann-Whitney Rank test also showed no significant difference between accuracy distributions across lighting conditions. These results indicate that there was no systematic advantage across all textures for either lighting condition, although some metamers may have been easier to spot under one lighting condition versus the other.

We next looked at whether characteristics of the textures predict metamer success, and so turned to the six texture descriptors described in Section 5.1. To account for both the correlated variation across repeated measures within participants and the missing data from accidental late responses, we performed a linear mixed-effects analysis [89] of the relationship between accuracy and descriptor. Fixed effects included each of the six descriptors and all pairwise interaction terms, with a random effect of subject. None of the descriptors showed a significant main effect, but we found two pairwise interactions significant at the $p < 0.01$ level: Contrast & Roughness ($\beta = -0.673$, SE = 0.214, $z(11375) = -3.142$, $p = 0.002$), and Coarseness & Roughness ($\beta = 0.589$, SE = 0.135, $z(11375) = 4.355$, $p \ll 0.001$).

Figure 15 shows the relationship between the roughness descriptor and both the contrast and coarseness descriptors relative to the average proportion correct for each texture. Visual inspection suggests that the significant interactions found in our statistical analysis are primarily driven by data points on the extreme ends of the distributions (upper-right for contrast and lower-left for coarseness, respectively). In particular, the most successful metamers were achieved by textures with high contrast and high roughness, while least successful metamers typically had low coarseness and low roughness. It is interesting that in the contrast/roughness interaction, in particular, high contrast (in the context of high roughness) is associated with successful metamers, given that high contrast in a visual stimulus is typically associated with higher visibility; one might expect that any artifacts in the metamer process would be more detectable. Perhaps high-contrast images, in the context of roughness, provide a strong overall structure, which, if maintained, serves to mask smaller scale and low-contrast variations that would otherwise be giveaways. This interpretation is consistent with low roughness and low coarseness as predictors of unsuccessful metamers, as these textures, although they may not necessarily have low contrast, may not have a strong overall structure capable of such masking.

In addition, Figure 16 shows a few examples of textures at the extreme ends of these distributions. Based on visual inspection, one notable difference between the hard-to-detect metamers on the left and the easy-to-detect metamers on the right is the very high degree of *both* local and global regularity in the textures on the left. Highly regular textures may lead to poor metamers, because if one detects even a small deviation from regularity, that provides a cue to distinguish that image from the original. Why, then, did our measure of regularity not

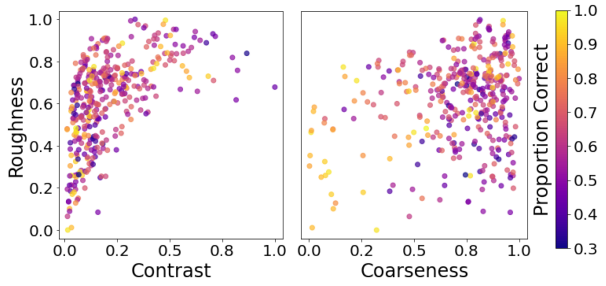

Fig. 15. Relationship between accuracy (proportion correct, color-coded) and texture descriptors for the two interactions we observed: Roughness vs. Contrast and Roughness vs. Coarseness. Darker colors indicate more successful metamers, while lighter colors less successful metamers. A high positive correlation with the x-axis variable, with no interaction, for example, would show a fade from dark to light along the x-axis but a random distribution of colors along the y-axis. However, a strong interaction would show a cluster of dark (or light) colors only in one corner of the plot. Here, roughness vs. contrast (left) shows primarily successful metamers in the top-right. Roughness vs. coarseness (right) shows primarily unsuccessful metamers in the bottom-left.

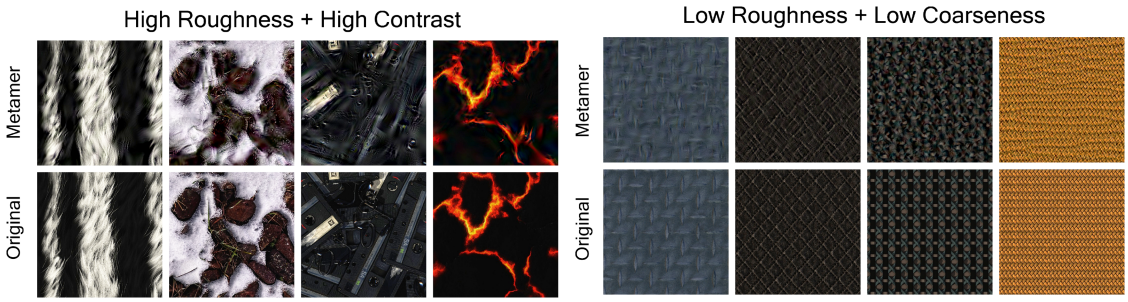

Fig. 16. Textures corresponding to interactions between the texture properties identified in Figure 15. The highest roughness combined with the highest contrast (left) are successful metamers that avoid detection. The lowest roughness combined with the lowest coarseness (right) are unsuccessful metamers that are detectable. Contrast enhanced for better viewing at a small scale. Best viewed electronically at full resolution.

distinguish between good and bad metamers? Our definition of regularity merely required textures to be regular at *any one scale*; any irregularity at another scale might reduce the cues distinguishing the original from metamer image, leading to a more successful metamer. Regardless, our primary finding is that textures with cross-scale regularity do not appear to make good metamers using our synthesis procedure. It is possible that that our model fails to account for some longer-range structure that the visual system is able to perceive, or this may simply be due to a failure of the synthesis procedure to fully converge in the case of textures with very low randomness. Fortunately, the flexibility of our model allows us to test many variations of statistics with ease, and we are continuing to explore better ways to synthesize successful metamers even for these particularly difficult textures.

7 CONCLUSIONS AND FUTURE WORK

We have reviewed the current state-of-the-art in modeling peripheral crowding: a summary statistic encoding model. The summary statistic nature of peripheral vision has implications for applications such as graphics, due to the significant loss of information in the periphery. The encoding model could provide a loss function, for example, for an efficient peripheral renderer or encoder. However, development of this model, while it has been well

tested, remains in progress. More work remains to be done to finalize the spatial summary statistics measured by the model, and there does not yet exist a temporal model that one can apply to video input. One of the main practical barriers to vision scientists making progress has been that previous computational approaches have led to very long run times on relatively small images. Another barrier has been the difficulty adding to or changing the model summary statistics. Both of these issues have led to difficulty testing new candidate statistics. We have described a different computational approach, based on the type of data-flow computation used in most deep learning tool-kits, that greatly improves the boundaries of what inputs can be processed and how quickly. The PyTorch code is available at https://github.com/ProgramofComputerGraphics/PooledStatisticsMetamers. Further, we conducted a psychophysical study to characterize the texture domains in which our version of the summary statistics model produces images difficult to distinguish from the original.

Our work can likely be applied immediately in domains such as texture synthesis, but ultimately the biggest payoffs, in animated systems, will need to account for temporal aspects of peripheral vision. Temporal aspects of peripheral crowding may or may not prove to be well-modeled by temporal or spatio-temporal summary statistics; early evidence suggests that crowding occurs across time as well as space, and that temporal crowding has similar characteristics to spatial crowding [90, 91]. The tools introduced in this article will enable vision science researchers to explore that space of possibilities through reducing the computation needed, and facilitating testing of new statistics. Graphics researchers might also be able to use spatial summary statistics in the current model in conjunction with empirical work on motion perception; such an approach has proven fruitful in non-peripheral contexts such as temporal anti-aliasing. We hope that an updated understanding of peripheral vision, in conjunction with computational techniques from graphics and machine vision, will lead to a productive area of collaboration between graphics researchers and vision scientists, as many other areas have in the past.

ACKNOWLEDGMENTS

We would like to thank our colleagues Ben Boudaoud and Josef Spjut for their help rendering the texture images used in the experiment, and Don Greenberg for his support throughout the project.

REFERENCES

[1] Anton S. Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. 2019. DeepFovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Trans. Graph.* 38, 6 (2019), 1–13.

[2] James D. Basinger, John M. Wilson, and Robert A. Fisher. 1982. *The Technical Contributions of the Tactical Combat Trainer Development Program.* Technical Report. Aeronatical Systems Division, Wright-Patterson AFB.

[3] G. Osterberg. 1935. *Topography of the Layer of Rods and Cones in the Human Retina.* A. Busck. Retrieved from https://books.google.com/books?id=DeDrSAAACAAJ.

[4] Jyrki Rovamo, Veijo Virsu, and Risto Näsänen. 1978. Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision. *Nature* 271, 5640 (1978), 54–56.

[5] Christoph Zetzsche and Erhardt Barth. 1990. Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vis. Res.* 30, 7 (1990), 1111–1117.

[6] Ruth Rosenholtz, Jie Huang, Alvin Raj, Benjamin J. Balas, and Livia Ilie. 2012. A summary statistic representation in peripheral vision explains visual search. *J. Vis.* 12, 4 (2012), 14–14.

[7] Xuetao Zhang, Jie Huang, Serap Yigit-Elliott, and Ruth Rosenholtz. 2015. Cube search, revisited. *J. Vis.* 15, 3 (2015), 9–9.

[8] Honghua Chang and Ruth Rosenholtz. 2016. Search performance is better predicted by tileability than presence of a unique basic feature. *J. Vis.* 16, 10 (2016), 13–13.

[9] Preeti Sareen, Krista A. Ehinger, and Jeremy M. Wolfe. 2016. CB Database: A change blindness database for objects in natural indoor scenes. *Behav. Res. Meth.* 48, 4 (2016), 1343–1348.

[10] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. 2011. Peripheral vision and pattern recognition: A review. *J. Vis.* 11, 5 (2011), 13–13.

[11] Ruth Rosenholtz. 2016. Capabilities and limitations of peripheral vision. *Ann. Rev. Vis. Sci.* 2 (2016), 437–457.

[12] Anne M. Treisman and Garry Gelade. 1980. A feature-integration theory of attention. *Cogn. Psychol.* 12, 1 (1980), 97–136.

[13] Mary C. Potter. 1976. Short-term conceptual memory for pictures. *J. Experim. Psychol.: Hum. Learn. Mem.* 2, 5 (1976), 509.

[14] Kenneth C. Scott-Brown, Mark R. Baker, and Harry S. Orbach. 2000. Comparison blindness. *Visual Cogn.* 7, 1–3 (2000), 253–267.
[15] Daniel J. Simons and Daniel T. Levin. 1997. Change blindness. *Trends Cogn. Sci.* 1, 7 (1997), 261–267.
[16] H. Bouma. 1970. Interaction effects in parafoveal letter recognition. *Nature* 266 (Apr. 1970), 177–178.
[17] Dennis M. Levi, Stanley A. Klein, and A. P. Aitsebaomo. 1985. Vernier acuity, crowding and cortical magnification. *Vis. Res.* 25, 7 (1985), 963–977.
[18] Wilhelm Korte. 1923. Über die Gestaltauffassung im indirekten Sehen. *Zeitschrift für Psychologie* 93 (1923), 17–82.
[19] Hans Strasburger. 2014. Dancing letters and ticks that buzz around aimlessly: On the origin of crowding. *Perception* 43, 9 (2014), 963–976.
[20] Dennis M. Levi and Stanley A. Klein. 1986. Sampling in spatial vision. *Nature* 320, 6060 (1986), 360–362.
[21] Patrick J. Bennett and Martin S. Banks. 1991. The effects of contrast, spatial scale, and orientation on foveal and peripheral phase discrimination. *Vis. Res.* 31, 10 (1991), 1759–1786.
[22] Ingo Rentschler and Bernhard Treutwein. 1985. Loss of spatial phase relationships in extrafoveal vision. *Nature* 313, 6000 (1985), 308–310.
[23] Jerome Y. Lettvin. 1976. On seeing sidelong. *The Sciences* 16, 4 (1976), 10–20.
[24] Jeremy Freeman and Eero P. Simoncelli. 2011. Metamers of the ventral stream. *Nat. Neurosci.* 14, 9 (2011), 1195–1201.
[25] Jan Koenderink, Whitman Richards, and Andrea J. van Doorn. 2012. Space-time disarray and visual awareness. *i-Perception* 3, 3 (2012), 159–165.
[26] Dennis M. Levi. 2008. Crowding–An essential bottleneck for object recognition: A mini-review. *Vis. Res.* 48, 5 (2008), 635–654.
[27] Denis G. Pelli and Katharine A. Tillman. 2008. The uncrowded window of object recognition. *Nat. Neurosci.* 11, 10 (2008), 1129–1135.
[28] Michael H. Herzog, Bilge Sayim, Vitaly Chicherov, and Mauro Manassi. 2015. Crowding, grouping, and object recognition: A matter of appearance. *J. Vis.* 15, 6 (2015), 5–5.
[29] Benjamin Balas, Lisa Nakano, and Ruth Rosenholtz. 2009. A summary-statistic representation in peripheral vision explains visual crowding. *J. Vis.* 9, 12 (2009), 13–13.
[30] Laura Parkes, Jennifer Lund, Alessandra Angelucci, Joshua A. Solomon, and Michael Morgan. 2001. Compulsory averaging of crowded orientation signals in human vision. *Nat. Neurosci.* 4, 7 (2001), 739–744.
[31] Denis G. Pelli, Melanie Palomares, and Najib J. Majaj. 2004. Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *J. Vis.* 4, 12 (2004), 12–12.
[32] Ruth Rosenholtz. 2011. What your visual system sees where you are not looking. In *Human Vision and Electronic Imaging XVI*, Vol. 7865. International Society for Optics and Photonics, 786510.
[33] Javier Portilla and Eero P. Simoncelli. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* 40, 1 (2000), 49–70.
[34] Ruth Rosenholtz. 2015. Texture perception. https://doi.org/10.1167/15.3.9
[35] Ruth Rosenholtz. 1999. A simple saliency model predicts a number of motion popout phenomena. *Vis. Res.* 39, 19 (1999), 3157–3163.
[36] Aude Oliva, Antonio Torralba, Monica S. Castelhano, and John M. Henderson. 2003. Top-down control of visual attention in object detection. In *Proceedings of the International Conference on Image Processing*. IEEE, I–253.
[37] Dashan Gao and Nuno Vasconcelos. 2007. Bottom-up saliency is a discriminant process. In *Proceedings of the IEEE 11th International Conference on Computer Vision*. IEEE, 1–6.
[38] Isamu Motoyoshi, Shin'ya Nishida, Lavanya Sharan, and Edward H. Adelson. 2007. Image statistics and the perception of surface qualities. *Nature* 447, 7141 (2007), 206–209.
[39] Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 3 (2001), 145–175.
[40] Shaiyan Keshvari and Ruth Rosenholtz. 2016. Pooling of continuous features provides a unifying account of crowding. *J. Vis.* 16, 3 (2016), 39–39.
[41] 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. *Front. Psychol.* 3 (2012), 13.
[42] Krista A. Ehinger and Ruth Rosenholtz. 2016. A general account of peripheral encoding also predicts scene perception performance. *J. Vis.* 16, 2 (2016), 13–13.
[43] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. Texture synthesis using convolutional neural networks. *arXiv preprint arXiv:1505.07376* (2015).
[44] Thomas S. A. Wallis, Christina M. Funke, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, and Matthias Bethge. 2017. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *J. Vis.* 17, 12 (2017), 5–5.
[45] Arturo Deza, Aditya Jonnalagadda, and Miguel Eckstein. 2017. Towards metamerism via foveated style transfer. *arXiv preprint arXiv:1705.10041* (2017).
[46] 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 Trans. Graph.* 40, 4 (2021), 1–14.
[47] W. T. Freeman and E. H. Adelson. 1991. The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.* 13, 9 (1991), 891–906. DOI: http://dx.doi.org/10.1109/34.93808

[48] Ruth Rosenholtz, Dian Yu, and Shaiyan Keshvari. 2019. Challenges to pooling models of crowding: Implications for visual mechanisms. *J. Vis.* 19, 7 (2019), 15–15.
[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035.
[50] Ming Ronnier Luo. 2015. Cielab. *Encyc. Color Sci. Technol.* (2015), 1–7.
[51] John Stommel. 2016. Compact Orange Pepper Plants Bear Upright, Pungent Fruit. (2016). Photo courtesy of USDA Agricultural Research Service.
[52] David H. Hubel and Torsten N. Wiesel. 1959. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* 148, 3 (1959), 574–591.
[53] David H. Hubel and Torsten N. Wiesel. 1965. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol.* 28, 2 (1965), 229–289. DOI: http://dx.doi.org/10.1152/jn.1965.28.2.229
[54] Friedrich Heitger, Lukas Rosenthaler, Rüdiger Von Der Heydt, Esther Peterhans, and Olaf Kübler. 1992. Simulation of neural contour mechanisms: From simple to end-stopped cells. *Vis. Res.* 32, 5 (1992), 963–981. DOI: http://dx.doi.org/10.1016/0042-6989(92)90039-L
[55] Bela Julesz. 1981. Textons, the elements of texture perception, and their interactions. *Nature* 290, 5802 (1981), 91–97.
[56] Christof Zetzsche, Erhardt Barth, and Bernhard Wegmann. 1993. *The Importance of Intrinsically Two-dimensional Image Features in Biological Vision and Picture Coding*. MIT Press, Cambridge, MA, 109–138.
[57] Sarah Rosen, Ramakrishna Chakravarthi, and Denis G. Pelli. 2014. The Bouma law of crowding, revised: Critical spacing is equal across parts, not objects. *J. Vis.* 14, 6 (2014).
[58] Thomas S. A. Wallis, Christina M. Funke, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, and Matthias Bethge. 2019. Image content is more important than Bouma's Law for scene metamers. *ELife* 8 (2019), e42512.
[59] Lance Williams. 1983. Pyramidal parametrics. In *Proceedings of the 10th Annual Conference on Computer Graphics and Interactive Techniques*. 1–11.
[60] 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 Trans. Graph.* 35, 6 (2016), 1–12.
[61] Benjamin J. Balas. 2006. Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vis. Res.* 46, 3 (2006), 299–309.
[62] Phil Brodatz. 1966. *Textures: A Photographic Album for Artists and Designers*. Dover Publications, New York.
[63] Shaiyan Keshvari and Maarten Wijntjes. 2016. Peripheral material perception. *J. Vis.* 16, 12 (2016), 641–641.
[64] Lavanya Sharan, Ruth Rosenholtz, and Edward Adelson. 2009. Material perception: What can you see in a brief glance? *J. Vis.* 9, 8 (2009), 784–784.
[65] João Paulo. 2021. 3D Textures Website. Retrieved from http://3dtextures.me.
[66] Lennart Demes. 2021. CC0 Textures Website. Retrieved from http://cc0textures.com.
[67] Andreas Siess. 2021. Pattern Panda Website. Retrieved from http://patternpanda.org.
[68] Frederic Hoffmann. 2021. Public Domain Textures Website. Retrieved from http://publicdomaintextures.com.
[69] Brian. 2021. Free PBR Website. Retrieved from http://freepbr.com.
[70] Julio Sillet. 2021. Textures Webpage. Retrieved from https://gumroad.com/juliosillet.
[71] Dorian Zgraggen. 2021. CG Bookcase Website. Retrieved from https://www.cgbookcase.com/textures.
[72] Ozgen Karagol Arslan and M. Tolga Arslan. 2021. Share Textures Website. Retrieved from https://www.sharetextures.com/.
[73] Chris Ebbinger. 2021. Pixel Furance Free Game Textures Website. Retrieved from https://textures.pixel-furnace.com/.
[74] 2021. Texture Can Website. Retrieved from https://www.texturecan.com.
[75] Erica Grenci. 2021. PBR Texture Website (Purchased Texture Files). Retrieved from https://pbrtexture.com.
[76] IMAGE PROMOTION ASSOCIATION. 2021. Sketch Up Texture Club Website. (2021). Retrieved from https://www.sketchuptextureclub.com/textures.
[77] Texturebox Game Technologies Inc. 2021. Texture Box Website. Retrieved from https://www.texturebox.com.
[78] Poly Haven. 2021. Texture Haven Website. Retrieved from https://texturehaven.com/textures/.
[79] 2021. Cadhatch Website. Retrieved from http://www.cadhatch.com.
[80] Morgan McGuire, Michael Mara, and Zander Majercik. 2017. The G3D Innovation Engine. Retrieved from https://casual-effects.com/g3d. https://casual-effects.com/g3d.
[81] Gaius Cornelius. 2017. Image of Clouds Directly Overhead. Taken on the Evening of the Summer Solstice 2017. Retrieved from https://commons.wikimedia.org/wiki/File:Summer_Solstice_Clouds_2017_A.jpg.
[82] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*.
[83] A. Ravishankar Rao and Gerald L. Lohse. 1996. Towards a texture naming system: Identifying relevant dimensions of texture. *Vis. Res.* 36, 11 (1996), 1649–1669.

[84] Fang Liu and Rosalind W. Picard. 1996. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 18, 7 (1996), 722–733.
[85] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. 1978. Textural features corresponding to visual perception. *IEEE Trans. Syst. Man Cyber.* 8, 6 (1978), 460–473.
[86] Claudio Cusano, Paolo Napoletano, and Raimondo Schettini. 2021. T1K+: A database for benchmarking color texture classification and retrieval methods. *Sensors* 21, 3 (2021), 1010.
[87] Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behav. Res. Meth.* 51, 1 (2019), 195–203.
[88] Henry B. Mann and Donald R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* 18, 1 (1947), 50–60.
[89] Mary J. Lindstrom and Douglas M. Bates. 1988. Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Amer. Statist. Assoc.* 83, 404 (1988), 1014–1022.
[90] Peter J. Bex and Steven C. Dakin. 2005. Spatial interference among moving targets. *Vis. Res.* 45, 11 (2005), 1385–1398.
[91] Alex O. Holcombe. 2009. Seeing slow and seeing fast: Two limits on perception. *Trends Cogn. Sci.* 13, 5 (2009), 216–221.

Received 16 August 2021; revised 26 July 2022; accepted 5 August 2022