Wasserstein Distortion: Unifying Fidelity and Realism

Yang Qiu, Aaron B. Wagner
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853 USA
{yq268,wagner}@cornell.edu

Johannes Ballé
Google Research
New York, NY 10011 USA
jballe@google.com

Lucas Theis
Google DeepMind
London, UK
theis@google.com

2024 58th Annual Conference on Information Sciences and Systems (CISS) | 979-8-3503-6929-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/CISS59072.2024.10480168

***Abstract*—We introduce a distortion measure for images, Wasserstein distortion, that simultaneously generalizes pixel-level fidelity on the one hand and realism or perceptual quality on the other. We discuss its metric properties. Pairs of images that are close under Wasserstein distortion illustrate its utility. In particular, we generate random images that have high fidelity to a reference image in one location of the image and smoothly transition to an independent realization as one moves away from this point. Wasserstein distortion represents a generalization and synthesis of prior work on texture generation, image realism and distortion, and models of the early human visual system, in the form of an optimizable metric in the mathematical sense.**

***Index Terms*—Information Theory, Realism, Texture Synthesis, Distortion-Realism Tradeoff, Distortion-Perception Tradeoff**

I. Introduction

Classical image compression algorithms are optimized to achieve high pixel-level fidelity between the source and the reconstruction. That is, one views images as vectors in Euclidean space and seeks to minimize the distance between the original and reproduction using metrics such as PSNR, SSIM, etc. While effective to a large extent, these objectives have long been known to introduce artifacts, such as blurriness, into the reconstructed image [1]. Similar artifacts arise in image denoising, deblurring, and super-resolution.

Recently, it has been observed that such artifacts can be reduced if one simultaneously maximizes the *realism*[1] of the reconstructed images. Specifically, one seeks to minimize the distance between some distribution induced by the reconstructed images and the corresponding distribution for natural images [2]. A reconstruction algorithm that ensures that these distributions are close will naturally be free of obvious artifacts; the two distributions cannot be close if one is supported on the space of crisp images and the other is supported on the space of blurry images. Image reconstruction under realism constraints has been a subject of intensive research of late, both of an experimental (e.g. [3]) and theoretical (e.g. [4]) nature.

Up to now, the dual objectives of fidelity and realism have been treated as distinct and even in tension (e.g. [2], [5]). Yet they represent two attempts to capture the same notion, namely the image quality perceived by a human observer. It is natural then to seek a simultaneous generalization of the two. Such a generalization could be more aligned with human perception than either objective alone, or even a linear combination of the two. The main contribution of this paper is one such generalization, *Wasserstein distortion*, which is grounded in models of the Human Visual System (HVS).

Realism objectives take several forms depending on how one induces a probability distribution from images: the distribution induced by the ensemble of full resolution images [6], a distribution over patches by selecting a patch at random from within a randomly selected image [7], or the distribution over patches within a given image induced by selecting a location at random and extracting the resulting patch [8]. Theoretical studies have tended to focus on the first approach while experimental studies have focused more on patches. We shall focus on the third approach because it lends itself more naturally to unification with fidelity: both depend only on the image under examination without reference to other images in the ensemble. That said, the proposed Wasserstein distortion can be extended naturally to videos and other sequences of images and in this way it generalizes the other notions of realism.

Our simultaneous generalization of fidelity and realism is based in theories of the HVS, as noted above; namely it resorts to computing *summary statistics* [9]–[11]. In particular, Freeman and Simoncelli [12] propose a model of the HVS focusing on the ventral stream. The visual field is divided into various receptive fields, and the ventral stream extracts information from each of them. The receptive fields grow with eccentricity, as depicted in Fig. 1. In the visual periphery, the receptive fields are large and only the response statistics are acquired. In the fovea, i.e., the center of gaze, the receptive field is assumed small enough that the statistics uniquely determine the image itself. See [12] for a complete description of the model. One virtue of this model is that it does not require separate theories of foveal and peripheral vision: the distinction between the two is simply the result of different receptive field sizes.

This unification of foveal and peripheral vision likewise suggests a way of unifying fidelity and realism objectives. For each location in an image, we compute the distribution of

The first two authors were supported by the US National Science Foundation under grant CCF-2306278 and a gift from Google.

[1]Realism is also referred to as *perceptual quality* by some authors.

Fig. 1. Receptive fields in the ventral stream grow with eccentricity.

features locally around that point using a weight function that decreases with increasing distance. The Wasserstein distance between the distributions computed for a particular location in two images measures the discrepancy between the images at that point. The overall distortion between the two images is then the sum of these Wasserstein distances across all locations. We call this *Wasserstein distortion.* If when constructing the distribution of features around a point, we use a strict notion of locality, i.e., a weight function that falls off quickly with increasing distance, then this reduces to a fidelity measure, akin to small receptive fields in the HVS model [12]. If we use a loose notion of locality, i.e., a weight function that falls off slowly with distance, then this reduces to a realism measure, akin to large receptive fields. Between the two is an intermediate regime with elements of both.

We propose the use of a one-parameter family of weight functions, where the parameter (σ) governs how strictly locality is defined. We find that to obtain good results requires careful selection of the family, especially its spectral properties. We prove that under a properly chosen weight function, Wasserstein distortion is a proper metric.

The balance of the paper is organized as follows. Section II consists of a mathematical description of Wasserstein distortion. Section III discusses metric properties of the distortion measure. Section IV contains our experimental results, specifically randomly generated images that are close to references under our distortion measure. An extended version of the paper with additional commentary and experimental results is available [13].

II. Definition of Wasserstein Distortion

We turn to defining Wasserstein distortion between a reference image, represented by a sequence $\mathbf{x} = \{x_n\}_{n=-\infty}^{\infty}$, and a reconstructed image, denoted by $\hat{\mathbf{x}} = \{\hat{x}_n\}_{n=-\infty}^{\infty}$. For notational simplicity, we shall consider 1-D sequences of infinite length, the 2-D case being a straightforward extension.

Let T denote the unit advance operation, i.e., if $\mathbf{x}' = T\mathbf{x}$ then

$$x'_n = x_{n+1}. \tag{1}$$

We denote the k-fold composition $T \circ T \circ \cdots \circ T$ by T^k. Let $\phi(\mathbf{x}) : \mathbb{R}^{\mathbb{Z}} \mapsto \mathbb{R}^d$ denote a vector of local features of $\{x_n\}_{n=-\infty}^{\infty}$ about $n = 0$. The simplest example is the coordinate map, $\phi(\mathbf{x}) = x_0$. More generally, $\phi(\cdot)$ can take the form of a convolution with a kernel $\alpha(\cdot)$

$$\phi(\mathbf{x}) = \sum_{k=-m}^{m} \alpha(k) \cdot x_k, \tag{2}$$

or, since ϕ may be vector-valued, it can take the form of a convolution with several kernels of the form in (2). There exist multiple choices of $\phi(\cdot)$: a steerable pyramid (e.g. [14]), convolution with a kernel as in (2) with random weights followed by a nonlinear activation function [15], a trained multi-layer convolutional neural network [16], etc.

Define the sequence $\mathbf{z}$ by

$$z_n = \phi(T^n \mathbf{x}) \tag{3}$$

and note that $z_n \in \mathbb{R}^d$ for each n. We view $\mathbf{z}$ as a representation of the image $\mathbf{x}$ in feature space.

Let $q_\sigma(k)$, $k \in \mathbb{Z}$, denote a family of probability mass functions (PMFs) over the integers, parameterized by $0 \le \sigma < \infty$, satisfying:

- **P.1** For any σ and k, $q_\sigma(k) = q_\sigma(-k)$;
- **P.2** For any σ and $k, k' \in \mathbb{Z}$ such that $|k| \le |k'|$, $q_\sigma(k) \ge q_\sigma(k')$;
- **P.3** If $\sigma = 0$, q_σ is the Kronecker delta function, i.e., $q_0(k) = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases}$;
- **P.4** For all k, $q_\sigma(k)$ is continuous in σ at $\sigma = 0$;
- **P.5** There exists $\epsilon > 0$ and K so that for all k such that $|k| \ge K$, $q_\sigma(k)$ is nondecreasing in σ over the range $[0, \epsilon]$; and
- **P.6** For any k, $\lim_{\sigma \to \infty} q_\sigma(k) = 0$.

We call $q_\sigma(\cdot)$ the *pooling PMF* and σ the *pooling width* or *pooling parameter*. One PMF satisfying **P.1**-**P.6** is the *two-sided geometric distribution*,

$$q_\sigma(k) = \begin{cases} \frac{e^{1/\sigma}-1}{e^{1/\sigma}+1} \cdot e^{-|k|/\sigma} & \text{if } \sigma > 0 \\ 1 & \text{if } \sigma = 0 \text{ and } k = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

In practice, the size of the pooling region, or equivalently σ, would vary across image. We define the *σ-map*: $\sigma(n)$ to allow σ to depend on n.

From the sequence $\mathbf{x}$, we define a sequence of probability measures $\mathbf{y}_\sigma = \{y_{n,\sigma(n)}\}_{n=-\infty}^{\infty}$ via

$$y_{n,\sigma(n)} = \sum_{k=-\infty}^{\infty} q_{\sigma(n)}(k) \delta_{z_{n+k}}, \tag{5}$$

where $\mathbf{z}$ is related to $\mathbf{x}$ through (3) and $\delta_\cdot$ denotes the Dirac delta measure. Each measure $y_{n,\sigma(n)}$ in the sequence represents the statistics of the features pooled across a region centered at n with effective width σ. Note that all measures in $\mathbf{y}$ share the same countable support set in $\mathbb{R}^d$; they differ only in the probability that they assign to the points in this set. See Fig. 2. Similarly, we define $\hat{\mathbf{x}} = \{\hat{x}_n\}_{n=-\infty}^{\infty}$, $\hat{\mathbf{z}} = \{\hat{z}_n\}_{n=-\infty}^{\infty}$, and $\hat{\mathbf{y}}_\sigma = \{\hat{y}_{n,\sigma(n)}\}_{n=-\infty}^{\infty}$ for the reconstructed image.

Let $d : \mathbb{R}^d \times \mathbb{R}^d \mapsto [0, \infty)$ denote an arbitrary distortion measure over the feature space. One natural choice is Euclidean distance

$$d(z, \hat{z}) = ||z - \hat{z}||_2, \tag{6}$$

although in general we do not even assume that d is a metric. We define the distortion between the reference and reconstructed images at location n to be

$$D_{n,\sigma(n)} = W_p^p \left(y_{n,\sigma(n)}, \hat{y}_{n,\sigma(n)} \right), \tag{7}$$

where W_p denotes the Wasserstein distance of order p [17, Def. 6.1][2]:

$$W_p(\rho, \hat{\rho}) = \inf_{Z \sim \rho, \hat{Z} \sim \hat{\rho}} \mathbb{E} \left[d^p(Z, \hat{Z}) \right]^{1/p}, \tag{8}$$

where ρ and $\hat{\rho}$ are probability measures on $\mathbb{R}^d$. The distortion over a block $\{-N, \ldots, N\}$ is defined as the spatial average

$$D = D(\mathbf{x}, \mathbf{x}') = \frac{1}{2N+1} \sum_{n=-N}^{N} D_{n,\sigma(n)}. \tag{9}$$

Wasserstein distance is widely employed due to its favorable theoretical properties. In practice one might adopt a proxy for (8) that is easier to compute. Following the approach used with Fréchet Inception Distance (FID) [18], one could replace (8) with

$$||\mu - \hat{\mu}||_2^2 + \mathrm{Tr}(C + \hat{C} - 2(\hat{C}^{1/2} C \hat{C}^{1/2})^{1/2}). \tag{10}$$

This is equivalent to W_p^p if we take $p = 2$, d to be Euclidean distance, and assume that ρ (resp. $\hat{\rho}$) is Gaussian with mean μ (resp. $\hat{\mu}$) and covariance matrix C (resp. $\hat{C}$). In our experiments, we simplify this even further by assuming that the features are uncorrelated,

$$\sum_{i=1}^{d} (\mu_i - \hat{\mu}_i)^2 + \left(\sqrt{V_i} - \sqrt{\hat{V}_i} \right)^2, \tag{11}$$

where μ_i and V_i are the mean and variance of the ith component under ρ and similarly for $\hat{\rho}$. This is justified when the feature set is overcomplete because the correlation between two features is likely to be captured by some third feature, as noted previously by [19]. Other possible proxies include sliced Wasserstein distance [20], Sinkhorn distance [21], Maximum Mean Discrepancy (MMD) [22], or the distance between Gram matrices [16].

The idea of measuring the discrepancy between images via the Wasserstein distance, or some proxy thereof, between distributions in feature space is not new (e.g. [19], [20]). As they are concerned with ergodic textures or image stylization, these applications effectively assume a form of spatial homogeneity, which corresponds to the regime of large pooling regions ($\sigma \to \infty$) in our formulation, and empirical distributions with equal weights over the pixels. That is, the pooling PMF in (5) is taken to be uniform over a large interval centered at zero (e.g., Eq. (1) of [23]). Our goal here is to lift fidelity and realism into a common framework by considering the full range of σ values, and we shall see next that for small or moderate values of σ, the uniform PMF is problematic.

[2]We refer to W_p as the Wasserstein *distance* even though it is not necessarily a metric if d is not a metric.

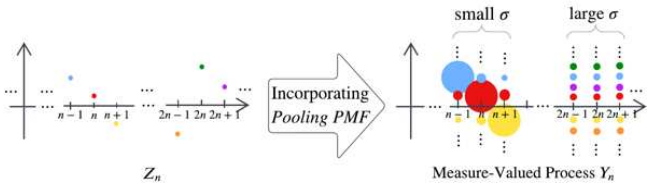

Fig. 2. A pictorial illustration of (5). In the right plot, the size of the disk indicates the probability mass and the vertical coordinate of the center of the disk indicates the value.

III. Metric Properties of Wasserstein Distortion

As $\sigma \to 0$, one can show that Wasserstein distortion converges to the conventional distortion between $\mathbf{z}$ and $\hat{\mathbf{z}}$ as measured by d raised to the p-th power [13]. If the source and reconstruction represent ergodic processes, then as $\sigma \to \infty$, Wasserstein distortion converges almost surely to the Wasserstein distance, again to the p-th power, between the marginal distributions of $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ [13]. In the $\sigma \to \infty$ regime, Wasserstein distortion will therefore not be a true metric in that certain pairs of distinct $\mathbf{x}$ and $\mathbf{x}'$ will have zero distortion, e.g., a pair of realizations drawn independently from the same ergodic process. Practically speaking, when σ is large, the Wasserstein distortion between two independent realizations of the same texture will be essentially zero. When σ is small, however, we want Wasserstein distortion to behave as a conventional distortion measure and as such it is desirable that it be a metric or a power thereof. In particular, we desire that it satisfy *positivity*, i.e., that $D(\mathbf{x}, \mathbf{x}') \geq 0$ with equality if and only if $\mathbf{x} = \mathbf{x}'$.

Whether Wasserstein distortion satisfies positivity at finite σ depends crucially on the choice of the pooling PMF. Consider, for example, the popular uniform PMF:

$$q_m(k) = \begin{cases} \frac{1}{2m+1} & \text{if } |k| \leq m \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

In this case Wasserstein distortion does not satisfy positivity, even over the feature space, for any m: a periodic signal with period $2m+1$ and its shifted variants would be distinct sequences with distortion zero. In practice, this means that Wasserstein distortion with a uniform pooling PMF is oblivious to certain blocking artifacts [13].

The problem lies with the spectrum of the pooling PMF. This is easiest to see in the case of MMD, for which the Wasserstein distortion reduces to the squared Euclidean distance between the convolution of the feature vectors with the pooling PMF. Thus if the pooling PMF has a spectral null, feature vectors that have all of their energy located at the null are indistinguishable from zero. Conversely, if the pooling PMF has no spectral nulls, then Wasserstein distortion is the p-th power of a metric, as we state next. For this theorem, we assume that $\mathbf{x}$ and $\mathbf{x}'$ (resp. $\mathbf{z}$ and $\mathbf{z}'$) are finite-length sequences, and the indexing in (5) is wraparound. For the proof, please see the full version [13].

Theorem III.1. *For any $0 \leq \sigma < \infty$, if d is a metric and $q_\sigma(\cdot)$ has no spectral nulls, then $D(\mathbf{z}, \mathbf{z}')^{1/p}$ is a metric. If, in addition, $\phi(\cdot)$ is invertible then $D(\mathbf{x}, \mathbf{x}')^{1/p}$ is also a metric.*

When σ is large, the PMF will be nearly flat over a wide range, so its spectrum will necessarily decay quickly. For small σ, the PMF is concentrated in time, so the spectrum can be made nearly flat in frequency if one chooses. Theoretically speaking, we need only to avoid PMFs with spectral nulls, such as the uniform distribution, to ensure positivity. Practically speaking, we desire pooling PMFs with a good *condition number*, meaning that the ratio of the maximum of the power spectrum to its minimum is small. In this vein, we note that the two-sided geometric PMF in (4) is well-conditioned, whereas the raised-cosine-type PMF used in [12, Eq. (9) with $t = 1/2$] has a condition number that is larger by almost four orders of magnitude for pooling regions around size 20. Note that papers in the literature that rely on uniform PMFs are focused on realism, i.e., the large σ regime, for which the presence of spectral nulls is less of a concern.

IV. Experiments

We validate Wasserstein distortion using the method espoused by [24], namely by taking an image of random pixels and iteratively modifying it to reduce its Wasserstein distortion to given a reference image. Following [16], we use as our feature map selected activations within the VGG-19 network with some modifications. We use the scalar Gaussianized Wasserstein distance in (11) as a computational proxy for (8). For the pooling PMF, we take the horizontal and vertical offsets to be i.i.d. according to the two-sided geometric distribution in (4), conditioned on landing within the boundaries of the image. We minimize the Wasserstein distortion between the reference and reconstructed images using the L-BFGS algorithm [25] with $4,000$ iterations and an early stopping criterion. For a detailed explanation, please refer to the full version [13].

A. Pinned Texture Synthesis

We consider texture images, with σ varying spatially over the image. Specifically, we set $\sigma = 0$ for pixels near the center; other pixels are assigned a σ proportional to their distance to the nearest pixel with $\sigma = 0$, with the proportionality constant chosen so that the outermost pixels have a σ that is comparable to the width of the image. The choice of having σ grow linearly with distance to the region of interest is supported by studies of the HVS. There are both physiological [26] and operational [12] evidences that the size of the receptive fields in the HVS grows linearly with eccentricity. If one seeks to produce images that are difficult for a human observer to easily distinguish, it is natural to match the pooling regions to the corresponding receptive fields when the gaze is focused on the $\sigma = 0$ region. Under this σ-map, Wasserstein distortion behaves like a fidelity measure in the center of the image and a realism measure along the edges, with an interpolation of the two in between. The results are shown in Fig. 3. The $\sigma = 0$ points have the effect of pinning the reconstruction to the original in the center, with a gradual transition to an independent realization at the edge.

B. Reproduction of Natural Images with Saliency Maps

We use the `SALICON` dataset [27] which provides a saliency map for each image that we use to produce a σ-map. Specifically, we set a saliency threshold above which pixels are declared high-salient. For such pixels we set $\sigma = 0$. For all other pixels σ is proportional to the distance to the nearest high-salient point, with the proportionality constant such that the farthest points should have a σ value on par with the width of the image.

The results are shown in Fig. 4. For images for which the non-salient regions are primarily textures, the reproductions are plausible replacements for the originals. In some other cases, the images appear to be plausible replacements if one focuses on high-salient regions, but not if one scrutinizes the entire image. This suggests that Wasserstein distortion can capture the discrepancy observed by a human viewer focused on high-salient regions.

It should be emphasized that the process of producing the reconstructions in Fig. 4 requires no pre-processing or manual labeling. In particular, it is not necessary to segment the image. Given a binarized saliency map, the σ-map can be constructed automatically using the above procedure, at which point the Wasserstein distortion is well defined.

Due to file size constraints, the images in this paper are compressed. For uncompressed images and additional results, please see the full version [13].

Acknowledgments

The authors wish to thank Eirikur Agustsson for calling their attention to [23].

References

[1] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[2] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6228–6237.

[3] O. Rippel and L. Bourdev, "Real-time adaptive image compression," in *International Conference on Machine Learning*. PMLR, 06–11 Aug 2017, pp. 2922–2930. [Online]. Available: https://proceedings.mlr.press/v70/rippel17a.html

[4] J. Klejsa, G. Zhang, M. Li, and W. B. Kleijn, "Multiple description distribution preserving quantization," *IEEE Transactions on Signal Processing*, vol. 61, no. 24, pp. 6410–6422, 2013.

[5] G. Zhang, J. Qian, J. Chen, and A. Khisti, "Universal rate-distortion-perception representations for lossy compression," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 517–11 529, 2021. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/5fde40544cff0001484ecae2466ce96e-Paper.pdf

[6] L. Theis and A. B. Wagner, "A coding theorem for the rate-distortion-perception function," in *Neural Compression: From Information Theory to Applications – Workshop @ ICLR 2021*, 2021. [Online]. Available: https://openreview.net/forum?id=BzUaLGtKecs

[7] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 221–231.

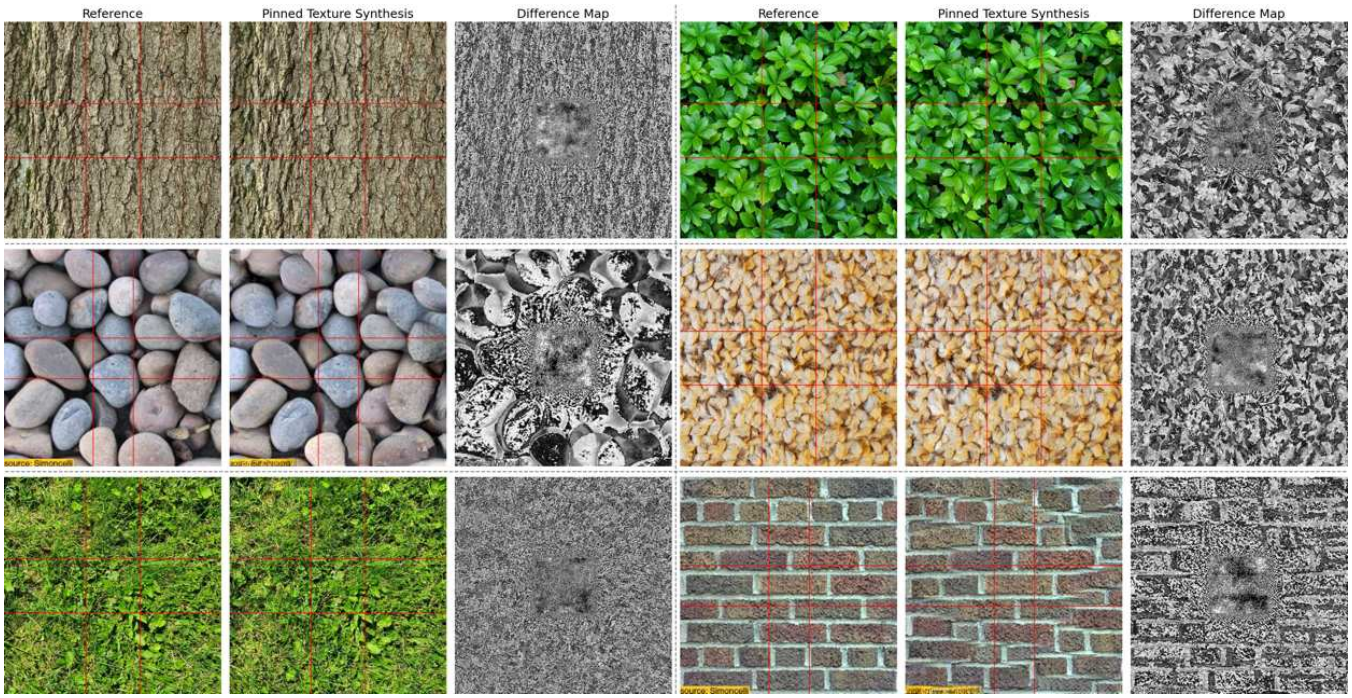

Fig. 3. Auxiliary lines indicate the square of $\sigma = 0$ points at the center. The reconstructions smoothly transition from pixel-level fidelity at the center to realism at the edges.

[8] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[9] B. Balas, L. Nakano, and R. Rosenholtz, "A summary-statistic representation in peripheral vision explains visual crowding," *Journal of Vision*, vol. 9, no. 12, pp. 13–13, 2009.

[10] R. Rosenholtz, "What your visual system sees where you are not looking," in *Human Vision and Electronic Imaging XVI*, vol. 7865. SPIE, 2011, pp. 343–356.

[11] R. Rosenholtz, J. Huang, A. Raj, B. J. Balas, and L. Ilie, "A summary statistic representation in peripheral vision explains visual search," *Journal of Vision*, vol. 12, no. 4, pp. 14–14, 2012.

[12] J. Freeman and E. P. Simoncelli, "Metamers of the ventral stream," *Nature Neuroscience*, vol. 14, no. 9, pp. 1195–1201, 2011.

[13] Y. Qiu, A. B. Wagner, J. Ballé, and L. Theis, "Wasserstein distortion: Unifying fidelity and realism," *arXiv:2310.03629*, 2023.

[14] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proceedings of the International Conference on Image Processing*, vol. 3, 1995, pp. 444–447 vol.3.

[15] I. Ustyuzhaninov, W. Brendel, L. Gatys, and M. Bethge, "What does it take to generate natural textures?" in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=BJhZeLsxx

[16] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/a5e00132373a7031000fd987a3c9f87b-Paper.pdf

[17] C. Villani, *Optimal Transport: Old and New*. Springer, 2009, vol. 338.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

[19] J. Vacher, A. Davila, A. Kohn, and R. Coen-Cagli, "Texture interpolation for probing visual perception," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 22 146–22 157. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/fba9d88164f3e2d9109ee770223212a0-Paper.pdf

[20] F. Pitié, A. Kokaram, and R. Dahyot, "n-dimensional probability density function transfer and its application to color transfer," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2, 2005, pp. 1434–1439.

[21] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in Neural Information Processing Systems*, vol. 26, 2013. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html

[22] A. J. Smola, A. Gretton, and K. Borgwardt, "Maximum mean discrepancy," in *13th International Conference, ICONIP*, 2006, pp. 3–6.

[23] E. Heitz, K. Vanhoey, T. Chambon, and L. Belcour, "A sliced Wasserstein loss for neural texture synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 9412–9420.

[24] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of full-reference image quality models for optimization of image processing systems," *International Journal of Computer Vision*, vol. 129, pp. 1258–1281, 2021.

[25] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 4, pp. 550–560, 1997.

[26] S. O. Dumoulin and B. A. Wandell, "Population receptive field estimates in human visual cortex," *Neuroimage*, vol. 39, no. 2, pp. 647–660, 2008.

[27] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1072–1080.

[28] G. Liu, Y. Gousseau, and G.-S. Xia, "Texture synthesis through convolutional neural networks and spectrum constraints," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 3234–3239.

[29] X. Snelgrove, "High-resolution multi-scale neural texture synthesis," in *SIGGRAPH Asia 2017 Technical Briefs*, 2017.

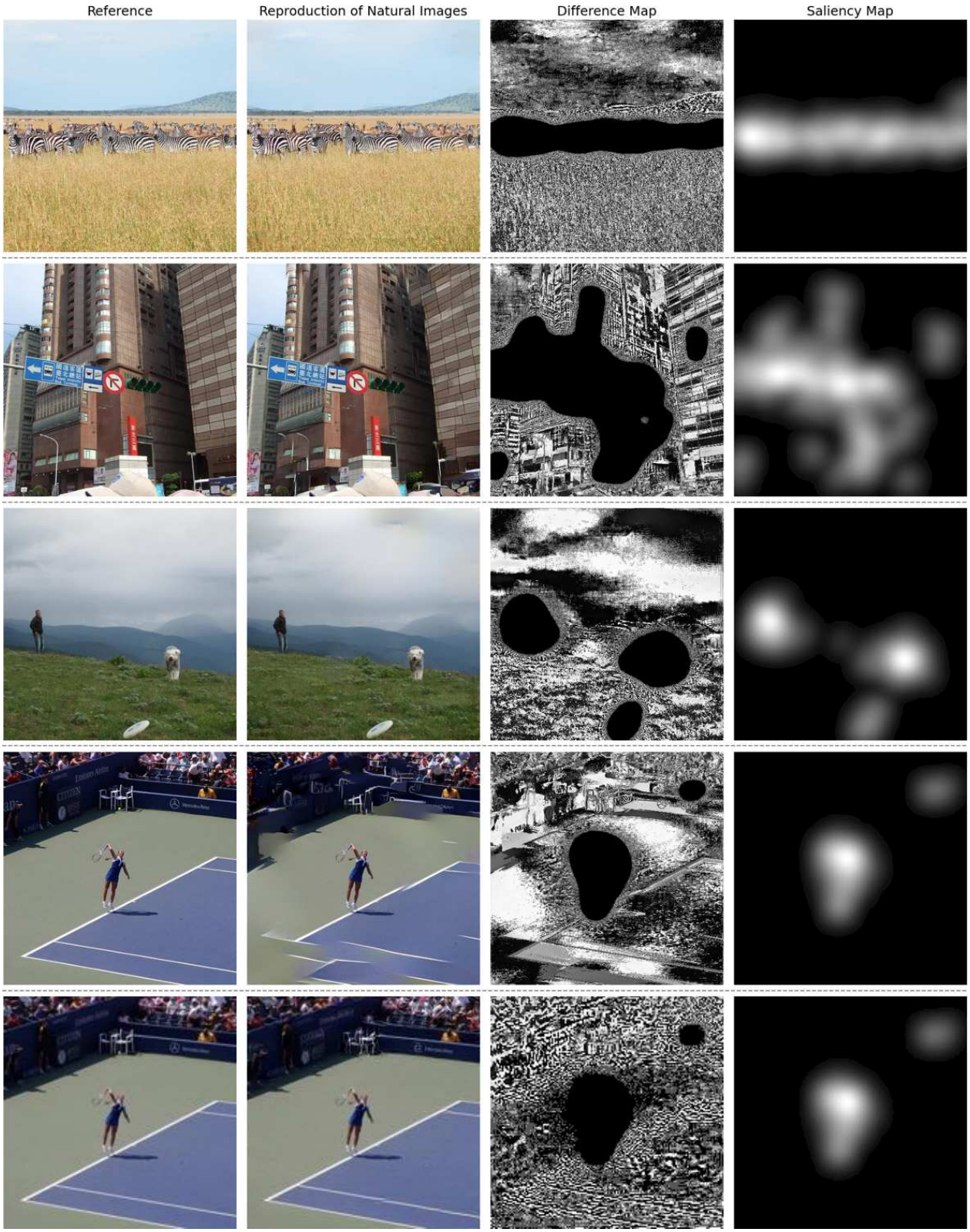

Fig. 4. For each row, the first image is the reference image and the second is the reproduction; the third is the difference between the two; and the fourth is the saliency map from SALICON before binarization. In the high saliency regions, the reconstruction exhibits pixel-level fidelity. Elsewhere, it exhibits realism or an interpolation of the two. Note that the goal of this experiment is not to reproduce images that withstand visual scrutiny in all regions, but to demonstrate how Wasserstein distortion becomes increasingly permissive to error towards the visual periphery, and that the errors that are permitted can be quite difficult to spot when viewing the salient regions at an appropriate distance. The misplaced foul lines in the fifth example are likely a manifestation of VGG-19's recognized difficulty with reproducing long linear features in textures (e.g. [28], [29]). This is evidenced through the last example, where the reference image has been downsampled so that VGG-19 better captures the long-range dependence. Compare with [12, Fig. 2].