Wasserstein Distortion with Intrinsic σ-Maps

Yang Qiu **Ziyuan Lin** **Aaron B. Wagner**
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853
{yq268,zl647,wagner}@cornell.edu

Abstract

Wasserstein distortion is a recently proposed family of distortion measures, controlled by a width parameter σ, that lifts fidelity and realism into a common framework. In previous implementations, calculating the Wasserstein distortion between two images relied on a companion saliency map or manual tuning to specify the width parameter σ for each location in the image. We introduce a novel scheme for automatically generating an σ-map from the image itself.

1 Introduction

Classical image compression algorithms aim to produce reconstructions that are close to the sources at the pixel-level. That is, one seeks to minimize a certain distance, e.g., MSE, PSNR, SSIM, etc., between the original and reproduction images. Reconstructions under these metrics are known to possess artifacts, especially blurriness [Wang and Bovik, 2009]. Similar artifacts also arise in other image-related tasks like denoising, deblurring, and super-resolution. Recently, *realism*[1] metrics, which measure distributional distances between the source and reconstruction, have been proposed to combat such artifacts [Blau and Michaeli, 2018]. A distribution over crisp images and another over distorted images would be heavily penalized by realism metrics, thus reducing the artifacts. Realism has been extensively studied recently, both experimentally (e.g., Agustsson et al. [2023], Iwai et al. [2024]) and theoretically (e.g., Hamdi et al. [2024], Serra et al. [2024], Salehkalaibar et al. [2024]).

While fidelity and realism constraints are sometimes seen as oppositional (e.g., Blau and Michaeli [2018], Zhang et al. [2021]), they both reflect a common purpose, i.e., to quantify image differences perceived by human. As such, attempts to merge both into one framework have been made; in particular, *Wasserstein distortion* was proposed recently in Qiu et al. [2024] (see also Qiu and Wagner [2024]). Wasserstein distortion measures the discrepancy between two images by calculating the divergence between local distributions derived from both images; the locality of the distributions is governed a parameter σ, called the *pooling width*, which can vary spatially over the image, and both fidelity and realism constraints are subsumed as extreme cases of the parameter. As the whole framework is inspired by a mathematical model of the early human vision system [Freeman and Simoncelli, 2011], one can attach a psychovisual interpretation to the σ value [Qiu et al., 2024]. A function that specifies σ for each pixel is called a *σ-map*. In previous implementations, the σ-map was either manually chosen or produced automatically using supplementary information such as a saliency map. We propose a novel algorithm to derive an *intrinsic* σ-map from a given source image. The idea is that for each pixel, we gradually increase the pooling width until the induced distribution begins to change. We show that using the new intrinsic σ-map in the tasks in Qiu et al. [2024] yields improved results.

The notion of a σ-map arises implicitly in the work of Freeman and Simoncelli [2011]. Given a reference image, they produce images whose statistics match those of the reference when pooled over

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

Workshop on Machine Learning and Compression @ NeurIPS 2024.

regions whose sizes vary spatially, being small in the center of the image and growing linearly as one moves toward the edges. Freeman and Simoncelli [2011] assume that the viewer will focus on the center of the image, and the larger pooling widths around the outside of the image take advantage of the fact that the peripheral vision only registers statistics pooled over regions. Thus a human viewer cannot distinguish between the two images as long as their focus remains in the center. In the framework of Wasserstein distortion, this corresponds to a particular choice of the σ-map that it is manually crafted based on properties of the human visual system and the assumption that only the center of the image is salient. Qiu et al. [2024] generalize the implicit σ-map of Freeman and Simoncelli [2011] by taking the σ value to be proportional to the distance to the nearest high saliancy region in the image, which requires access to a separate saliancy map. The σ-maps that we derive in this work, in contrast, are automatically generated from the statistics of the reference image itself. Our method bears some resemblance to *region growing* methods in image segmentation [Haralick and Shapiro, 1985], although a σ-map does not yield a segmentation of the image and vice versa.

The balance of the paper is organized as follows. Section 2 briefly reviews the definition of Wasserstein distortion and provides necessary definitions for the later parts. Section 3 provides the detailed scheme and explanation for the σ-map generation algorithm. Section 4 discusses technical aspects and future applications of the intrinsic σ-map.

2 Wasserstein Distortion

We give a brief overview for Wasserstein distortion, as introduced in Qiu et al. [2024].

Let $\mathbf{X} = \{X_{m,n}\}_{m=1,n=1}^{M,N}$ be a 2-D stochastic process that represents the source of interest, with realizations denoted by $\mathbf{x} = \{x_{m,n}\}_{m=1,n=1}^{M,N}$. In this work, we view the source as an image.

Let $\mathbf{Z} = \boldsymbol{\phi}(\mathbf{x})$ denote a tensor of local features of $\mathbf{x}$. With a slight abuse of notation, we also index $\mathbf{Z}$ by $Z_{m,n}$, though $\mathbf{Z}$ and $\mathbf{X}$ do not necessarily have the same block length. We also assume that for each (m,n), $z_{m,n}$ is a scalar. Like Qiu et al. [2024], we take $\boldsymbol{\phi}$ to be the output of selected layers of the VGG-19 network [Simonyan and Zisserman, 2015], although the framework does not require this.

Let $q_{m,n,\sigma}(k,l)$, $k = 1,2,\dots,M$, $l = 1,2,\dots,N$, denote a family of probability mass functions (PMFs) parameterized by a width parameter $0 \le \sigma < \infty$, symmetric about (m,n). We call $q_{m,n,\sigma}(\cdot,\cdot)$ the *pooling PMF* and σ the *pooling width*. Note that the definition of Wasserstein distortion is agnostic to the choice of the pooling PMF, subject to certain conditions (cf. Qiu et al. [2024, Section II]); in this work, we use a discretized Gaussian distribution with variance σ^2, truncated to the range of $\mathbf{Z}$:

$$q_{m,n,\sigma}(k,l) \propto \exp\left(-\frac{(k-m)^2+(l-n)^2}{2\sigma^2}\right), k = 1,2,\dots,M, l = 1,2,\dots,N. \tag{1}$$

Notice that when $\sigma = 0$, the PMF $q_{m,n,0}(k,l)$ becomes a Kronecker Delta at (m,n); and when $\sigma \to \infty$, the PMF converges to a uniform distribution over all pixels.

Given a realization $\mathbf{x}$, a location (m,n) where $m = 1,2,\dots,M$ and $n = 1,2,\dots,N$, and a pooling width σ, we define the probability measure

$$y_{m,n,\sigma} = \sum_{k=1}^{M}\sum_{l=1}^{N} q_{m,n,\sigma}(k,l)\delta_{z_{k,l}}, \tag{2}$$

where $\delta_\cdot$ denotes the Dirac delta function. Then $Y_{m,n,\sigma}$ is a random measure. Each realization $y_{m,n,\sigma}$ represents the statistics of the features pooled across a particular *pooling region* centered at (m,n) with width σ. Notice the two extreme cases: when $\sigma = 0$, the measure $y_{m,n,0}$ is simply a point mass on the pixel value of (m,n), and when $\sigma \to \infty$, the measure becomes a regular uniform empirical measure over the whole image. In practice, we might wish to use different values of σ for different locations n. We call the function $\sigma(m,n)$ that specifies σ for each (m,n) the *σ-map*.

Similarly, we can define $\hat{\mathbf{x}} = \{\hat{x}_{m,n}\}_{m=1,n=1}^{M,N}$, $\hat{\mathbf{z}} = \{\hat{z}_{m,n}\}_{m=1,n=1}^{M,N}$, $\hat{\mathbf{y}} = \{\hat{y}_{m,n,\sigma}\}_{m=1,n=1}^{M,N}$, etc., for the reconstruction process.

Consider any divergence between distributions $\mathcal{D}(\rho, \rho')$ over Euclidean space of a given dimension. The the Wasserstein distortion at location (m, n) is defined to be

$$D_{m,n,\sigma(m,n)} = \mathcal{D}\left(y_{m,n,\sigma(m,n)}, \hat{y}_{m,n,\sigma(m,n)}\right). \tag{3}$$

The Wasserstein distortion D between two images is defined as the spatial average

$$D = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} D_{m,n,\sigma(m,n)}. \tag{4}$$

Wasserstein distortion bridges the gap between fidelity and realism measures in that when $\sigma(m, n) = 0$, both distributions are point masses, and Wasserstein distortion reduces to a fidelity constraint; when $\sigma(m, n)$ is large, both distributions tend to uniform empirical distribution over the image, and Wasserstein distortion reduces to a realism measure.

For computational reasons, in this work we choose $\mathcal{D}$ to be the Fréchet Inception Distance (FID) [Heusel et al., 2017]; namely, for distributions ρ and ρ', their FID is

$$FID(\rho, \rho') = (\mu_\rho - \mu_{\rho'})^2 + (\sigma_\rho - \sigma_{\rho'})^2, \tag{5}$$

where μ_ρ, σ_ρ ($\mu_{\rho'}, \sigma_{\rho'}$, resp.) are the mean and standard deviation under distribution ρ (ρ', resp.).

3 Intrinsic σ-Map

In the definition above, for each location (m, n), we need to first specify the pooling width $\sigma(m, n)$ in order to calculate the Wasserstein distortion $D_{m,n,\sigma(m,n)}$. We propose a novel scheme to derive a σ-map from the local statistics of the image itself. As in image segmentation, we view the image as a collection of disjoint textures. Our goal would be, for each pixel location (m, n), to find the largest possible pooling width σ such that the corresponding empirical measure $y_{m,n,\sigma}$ does not penetrate into any neighboring texture. Our algorithm is developed based on the following observation: fix a location (m, n), consider two pooling widths σ_1 and σ_2, and their corresponding empirical measures y_{m,n,σ_1} and y_{m,n,σ_2}, respectively. We can measure the distortion between these two measures; if both y_{m,n,σ_1} and y_{m,n,σ_2} lie in the same texture, their distributional distance should be small; if, on the contrary, the larger pooling region penetrates into a neighboring texture while the smaller does not, their distributional distance should be large. Thus, we seek to find the best σ by computing the distortion between the same location in the same image for different σ values.

We first specify possible choices of σ values, sorted in ascending order. We then compute the distortion between pairs of empirical measures with consecutive σ's at a given location, and examine the list of distortion values. The algorithm divides pixel locations into two categories, high contrast and low contrast, depending on when the largest distortion occurs. A large discrepancy occurring between two small σ values indicates that the pixel is high contrast. Both local structural difference (e.g., different parts of human faces) and local disturbance (e.g., different blades of grass) could contribute to the distortion for pairs of small σ values. On the other hand, if the largest discrepancy occurs between two large values of σ, then we declare that the location is low contrast.

Consider the examples in Figs. 1-2. A natural proposition for finding the best σ value would be to find the pair of consecutive σ values for which the distortion is the largest. In practice, we found that this maximum occurs not when the pooling region first encounters the closest distinct texture but when it first penetrates the most distinct texture (Fig. 2). The maximum can also occur not when the pooling region first penetrates another texture but when the length of the boundary of the pooling region contained in the other texture is largest. We therefore find the first distortion that exceeds a fixed ratio times the peak distortion value. For high contrast peak, to eliminate local disturbances, we discount the peak by a fixed factor. We also exclude the early distortion values for the same reason. We then take the smaller σ in the pair, and set the σ value at the pixel location as the σ value a few prior to that in our list, again to mitigate the effect of other intruding textures.

We give a concise mathematical definition here. Consider the Wasserstein distortion between an image and itself; i.e., let the source $\mathbf{x}$ and the reconstruction $\hat{\mathbf{x}}$ be the same image. Fix a pixel location (m, n). Define a sequence of possible choices of σ-values, $\Sigma = \sigma_1, \ldots, \sigma_S$, where $\sigma_i < \sigma_j$ for all $i < j$. We calculate

$$D_{m,n,\sigma_i,\sigma_{i+1}} = \mathcal{D}\left(y_{m,n,\sigma_i}, \hat{y}_{m,n,\sigma_{i+1}}\right) =: d_i \tag{6}$$

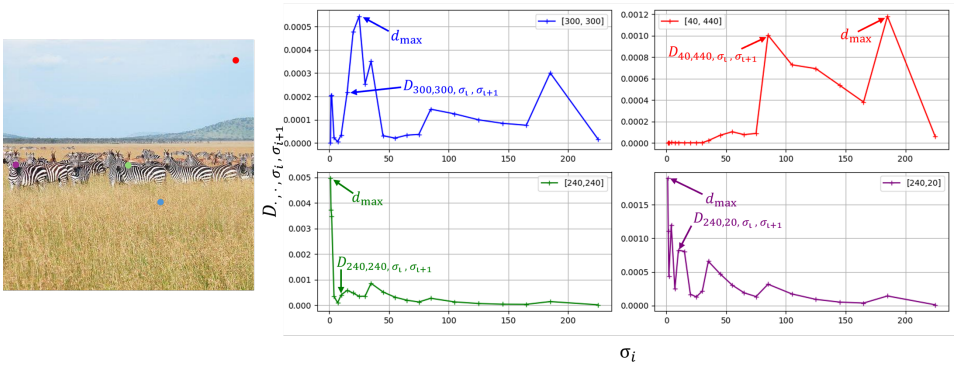

Figure 1: Examples of the distortion list. The source image is on the left, and the four distortion curves on the right correspond to the four pixel locations coded by color. The two distortion curves on the first row are low contrast, and the two on the second row are high contrast. We marked the peak value $d_{\max}$ and the first distortion that exceeds the relevant threshold.

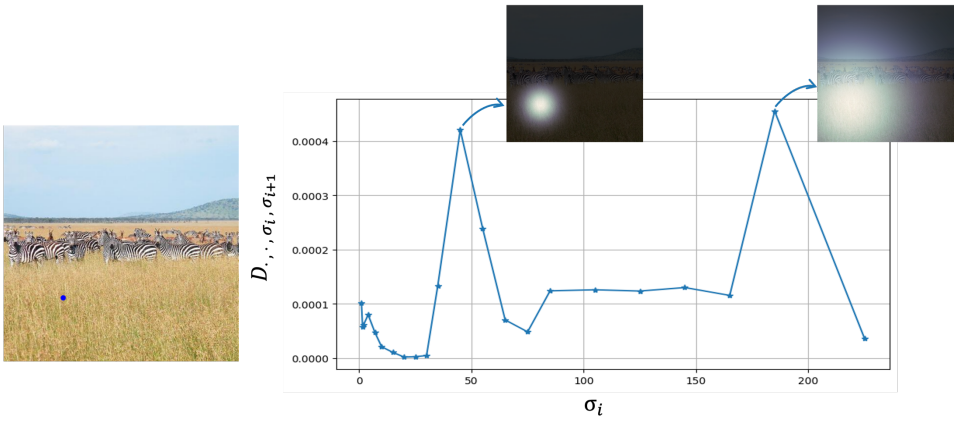

Figure 2: An example of the distortion list versus the σ list. The image on the left is the source image, and the curve represents the distortion list at pixel location (350,120). The right two images illustrate the empirical measure $y_{350,120,\sigma}$ for two different σ values, 45 and 185, respectively, where the luminance on each pixel indicates the probability of the corresponding pixel value. We see that while the distortion between the measures for $\sigma = 45$ and 55 is not the highest peak, this spike captures when the nearest neighboring different texture, namely the zebras, enters the measure, as opposed to the right peak where the most distinct texture, namely the sky, enters the measure.

for $i = 1, 2, \ldots, S-1$. Write the list of distortion values as $\mathbf{d} = d_1, \ldots, d_{S-1}$, and denote the maximum value in $\mathbf{d}$ as $d_{\max}$. Define an index threshold k. If the maximum occurs within the first k values, we declare the pixel location (m, n) as high contrast, and discount $d_{\max}$ by a fixed factor f and set $\bar{d} = d_{\max} \times f$. Otherwise, we declare the location low contrast and set $\bar{d} = d_{\max}$. We then define the threshold ratios τ. We remove the first k distortions, then compare the remaining distortion values $d_{k+1}, d_{k+2}, \ldots, d_{S-1}$ to the threshold $\tau \times \bar{d}$. We find the index ι such that d_ι is the first distortion value that exceeds $\tau \times \bar{d}$. We set the σ value of pixel location (m, n) to $\sigma_{\iota-\varsigma}$ if ι exists, where ς is the countback; otherwise, we set σ to 0.

The algorithm is summarized in Algorithm 1.

3.1 Metamer Reconstruction with Intrinsic σ-Map

We show the intrinsic σ-maps from our algorithm in Figures 3, 4, and Figures 5-8 in Appendix A. We set $\Sigma = \{1, 1.5, 2, 4, 7, 10, 15, 20, 25, 30, 35, 45, 55, 65, 75, 85, 105, 125, 145, 165, 185, 225\}$, $\tau = 0.35$, $f = 3/7$, $k = 5$, and $\varsigma = 1$. We implemented our algorithm on an Nvidia 3090 GPU with CUDA 11.8, Python 3.10 and TensorFlow 2.11. The average time to calculate an intrinsic σ-map for

Algorithm 1 Intrinsic σ-map Generation

Define list of possible σ-values $\Sigma = \{\sigma_1, \ldots, \sigma_S\}$, index threshold k, discount factor f, threshold ratio τ, and countback ς
Fix a pixel location (m, n), calculate $\mathbf{d} = \{d_1, d_2, \ldots, d_{S-1}\}$, and find $d_{\max} = \max \mathbf{d}$
if $d_{\max} \in \{d_1, d_2, \ldots, d_k\}$ **then**
    $\bar{d} = f \times d_{\max}$
else
    $\bar{d} = d_{\max}$
end if
Find $\iota := \min\{i \in \{k+1, k+2, \ldots, S-1\} : d_i \geq \tau \times \bar{d}\}$
if ι exists **then**
    $\sigma = \sigma_{\iota - \varsigma}$
else
    $\sigma = 0$
end if

a 480×480 image is 7 seconds. To illustrate the usage of the intrinsic σ-maps, we repeat the exact metamer reproduction experiment that appeared in Qiu et al. [2024, Experiment 4], and compared some results to the previous reconstructions there with the accompanying saliency maps. Metamer reconstruction has been found useful in exploring the properties of distortion measures [Ding et al., 2020], thus we use that to illustrate the properties of our intrinsic σ-map. We see that reproductions generated using the intrinsic σ-maps are more faithful to the original.

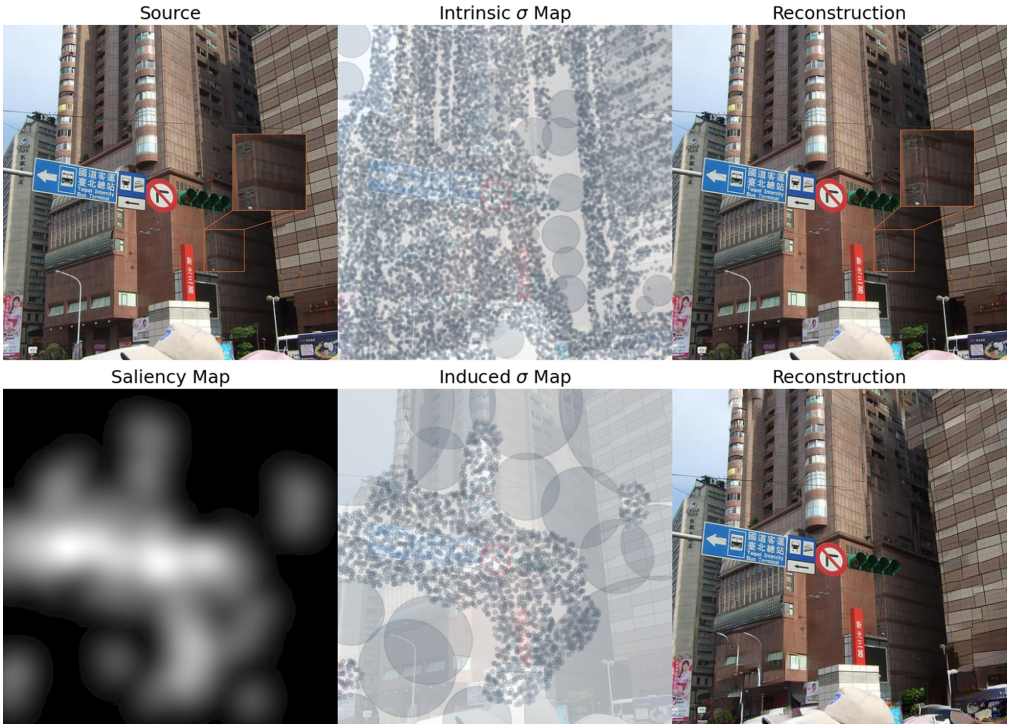

Figure 3: A comparison of our intrinsic σ-map versus the σ-map derived in Qiu et al. [2024], and their corresponding reconstructions. The first row consists of the source image, our intrinsic σ-map, and the corresponding metamer reconstruction; the second row is the saliency map from SALICON dataset [Jiang et al., 2015], the σ-map derived from that [Qiu et al., 2024], and the corresponding reconstruction. For the reconstruction, our goal is to produce a pixel-perfect reconstruction in the non-textural regions and produce a independent realization of the textures. We added insets in the source and intrinsic σ-map reconstruction to highlight differences in the textural areas. We see that our intrinsic σ-maps are better at identifying textural and non-textural areas.

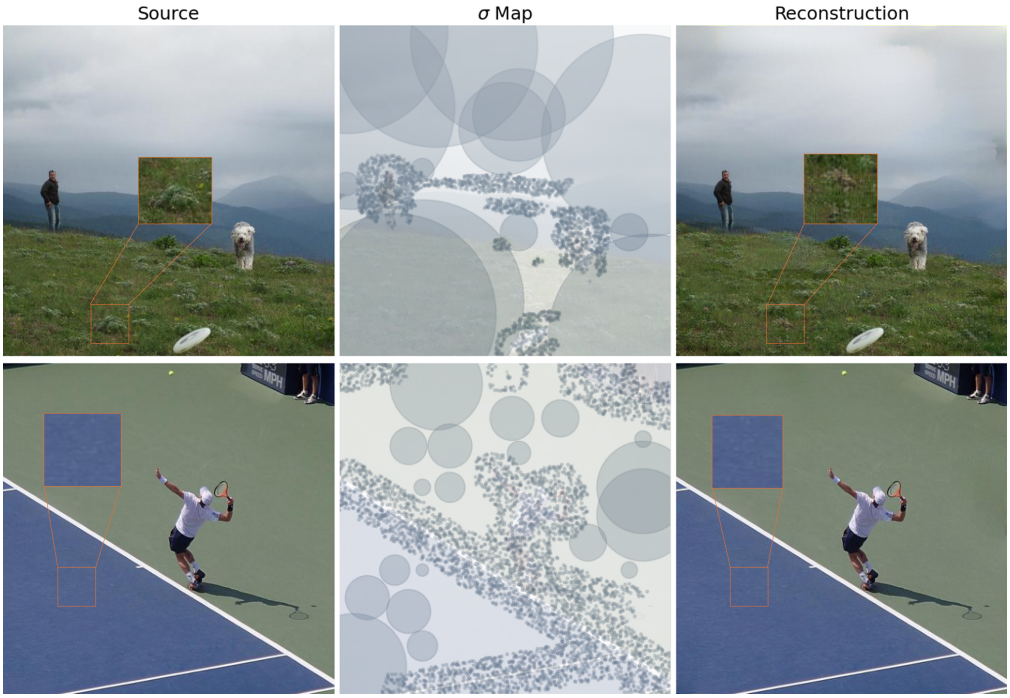

Figure 4: More samples. Each row consists of the source image, the intrinsic σ-map, and the corresponding reconstruction image. Difference between the source and reconstruction in the textural area is highlighted.

4 Conclusion and Discussion

We proposed an algorithm to derive a σ-map for an arbitrary image for use in Wasserstein distortion. Such a σ-map specifies a width parameter σ for each pixel location on the image. The parameter σ approximately determines a radius such that a disc centered at the pixel with the radius includes only one texture. Our σ-maps are validated via the metamer reproduction task.

We wish to point out that our choices of the parameters are independent of the resolution of the image; on various scaled-up/down versions of the images, the σ-maps produced under this set of parameters are similar. While the choice of Σ is arbitrary, we note that if the Σ list becomes too dense, the differences in neighboring empirical distributions are too small regardless of whether a new texture has landed in the range, which hurts the algorithm.

Appendix B provides some examples for which the algorithm does not find the most reasonable σ-map. For images consisting of a single texture, ideally σ would be large for the entire image. Instead, it sets σ to be zero everywhere since it is not able to identify a boundary where two textures meet (Fig. 9). Another challenge concerns text. Blocks of text resemble textures in important ways; but high fidelity is necessary for the text to be legible, so they should be treated as non-textural. An image consisting of nothing but text is treated as non-textural simply because of the reason noted above. We can create adversarial examples by surrounding a block of text by another texture; then the text itself will be treated as a texture, resulting in large σ-values and illegible reconstruction (Fig. 10).

Other uses of the σ-map are yet to be fully explored. A method for automatically generating σ-maps would enable Wasserstein distortion to be applied to compression, denoising [Elad et al., 2023], data embedding [Tao et al., 2014], etc. For instance, in image compression, minimizing the data rate subject to a constraint on the Wasserstein distortion, with the σ-map as generated here, could save bits by only encoding the statistics of large textural regions while still ensuring high pixel-level fidelity in the non-textural regions. Denoising, watermarking, and other image processing tasks that require a full reference image distortion measure could likewise benefit from using Wasserstein distortion with intrinsic σ-maps.

Acknowledgments and Disclosure of Funding

The authors would like to thank Jona Ballé for helpful discussions. This research was supported by the US National Science Foundation under grant CCF-2306278 and a gift from Google.

References

Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer. Multi-realism image compression with a conditional generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22324–22333, 2023.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. doi: 10.1109/cvpr.2018.00652.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

Michael Elad, Bahjat Kawar, and Gregory Vaksman. Image denoising: The deep learning revolution and beyond—a survey paper. *SIAM Journal on Imaging Sciences*, 16(3):1594–1654, 2023.

Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14 (9):1195–1201, 2011.

Yassine Hamdi, Aaron B. Wagner, and Deniz Gündüz. The rate-distortion-perception trade-off: the role of private randomness. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 1083–1088, 2024. doi: 10.1109/ISIT57864.2024.10619317.

Robert M Haralick and Linda G Shapiro. Image segmentation techniques. *Computer Vision, Graphics, and Image Processing*, 29(1):100–132, 1985.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Shoma Iwai, Tomo Miyazaki, and Shinichiro Omachi. Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2900–2909, 2024.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1072–1080, 2015.

Yang Qiu and Aaron B. Wagner. Low-rate, low-distortion compression with Wasserstein distortion. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 855–860, 2024. doi: 10.1109/ISIT57864.2024.10619196.

Yang Qiu, Aaron B. Wagner, Johannes Ballé, and Lucas Theis. Wasserstein distortion: Unifying fidelity and realism. In *2024 58th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, 2024. doi: 10.1109/CISS59072.2024.10480168.

Sadaf Salehkalaibar, Jun Chen, Ashish Khisti, and Wei Yu. Rate-distortion-perception tradeoff for lossy compression using conditional perception measure. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 1071–1076. IEEE, 2024.

Giuseppe Serra, Photios A Stavrou, and Marios Kountouris. On the computation of the gaussian rate-distortion-perception function. *IEEE Journal on Selected Areas in Information Theory*, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*, pages 1–14, 2015.

Hai Tao, Li Chongmin, Jasni Mohamad Zain, and Ahmed N Abdalla. Robust image watermarking theories and techniques: A review. *Journal of Applied Research and Technology*, 12(1):122–138, 2014.

Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009. doi: 10.1109/MSP.2008.930649.

George Zhang, Jingjing Qian, Jun Chen, and Ashish Khisti. Universal rate-distortion-perception representations for lossy compression. *Advances in Neural Information Processing Systems*, 34: 11517–11529, 2021.

A Additional Experimental Results for Section 3

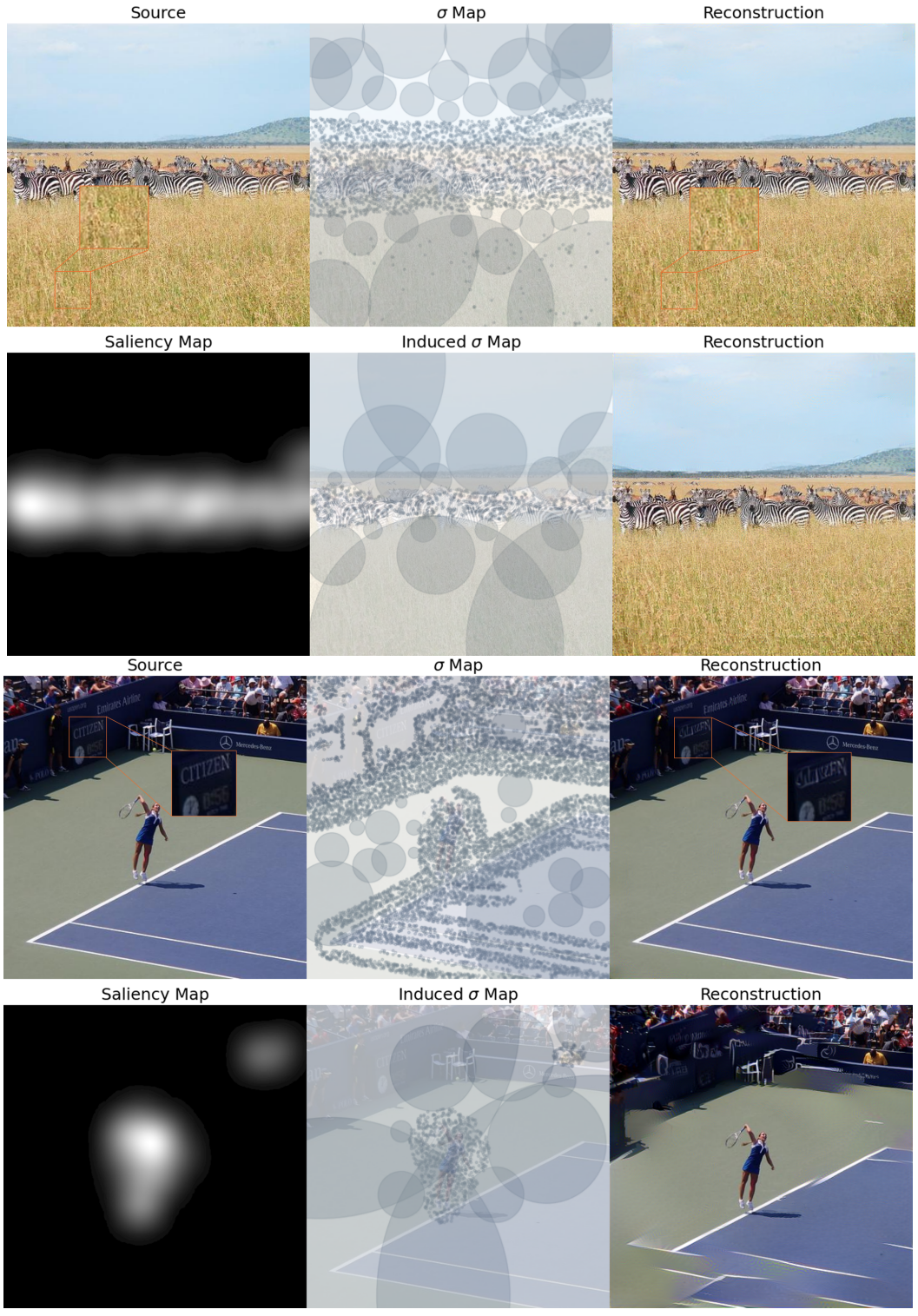

Figure 5: More results and their comparison to the previous saliency map-related results. The difference between the source and reconstruction in the textural area is highlighted.

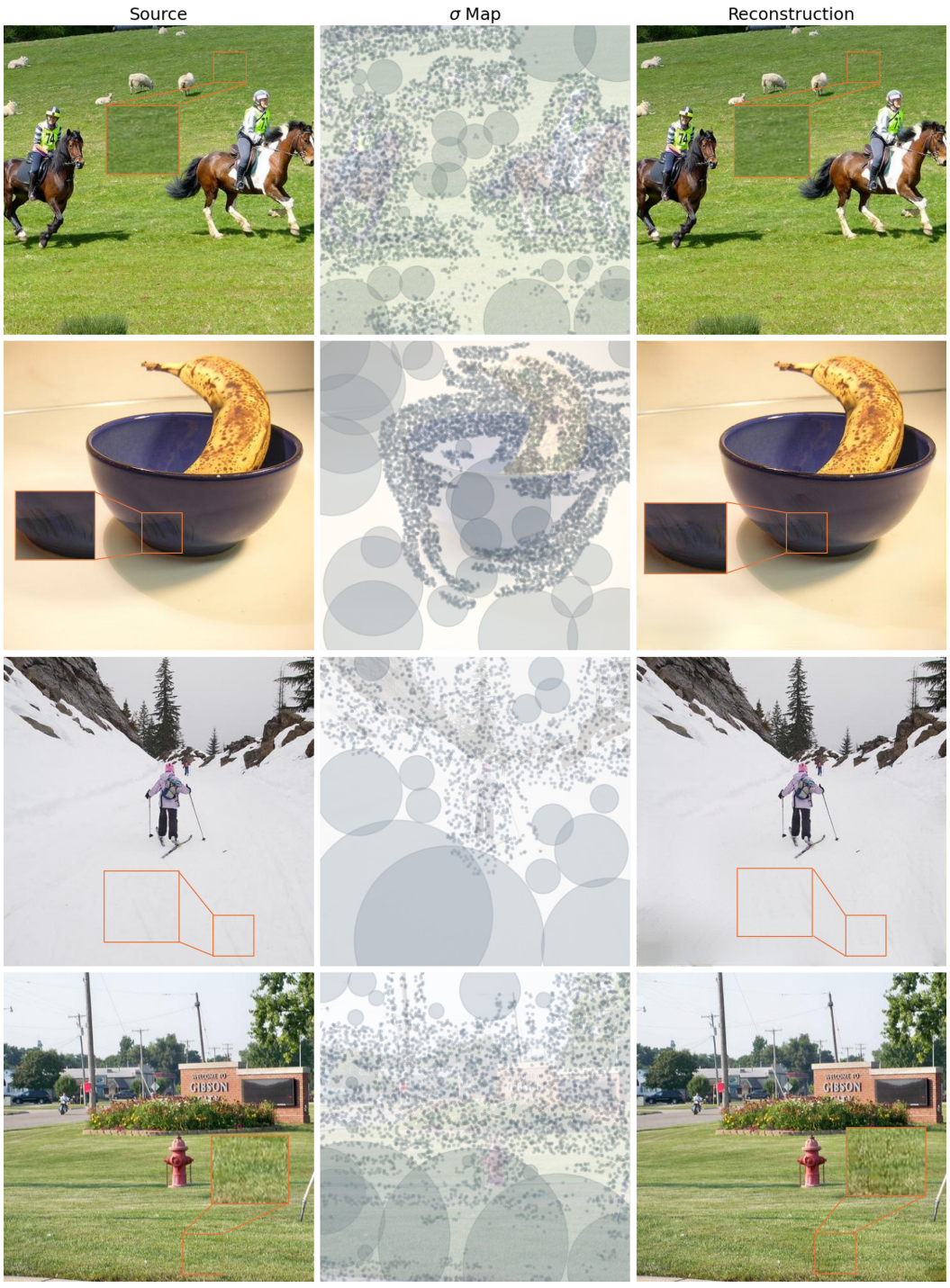

Figure 6: Additional results. The difference between the source and reconstruction in the textural area is highlighted.

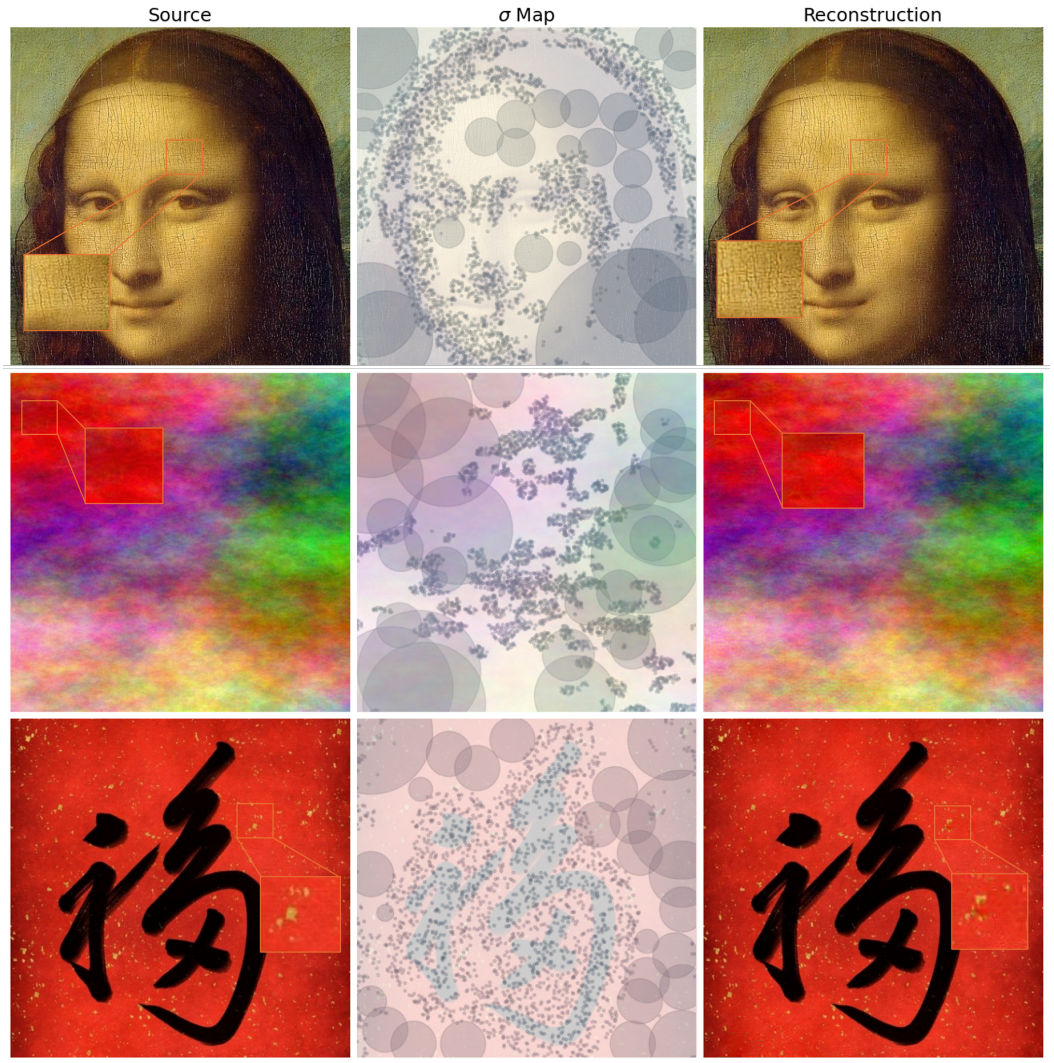

Figure 7: Additional results. All source images were drawn from datasets without accompanying saliency regions, and thus cannot be handled using the method of Qiu et al. [2024]. The difference between the source and reconstruction in the textural area is highlighted.

Figure 8: Additional results. All source images were drawn from datasets without accompanying saliency regions, and thus cannot be handled using the method of Qiu et al. [2024]. These images consist of mostly non-textural areas, and the source and reconstruction are largely identical, as desired.

B Negative Examples

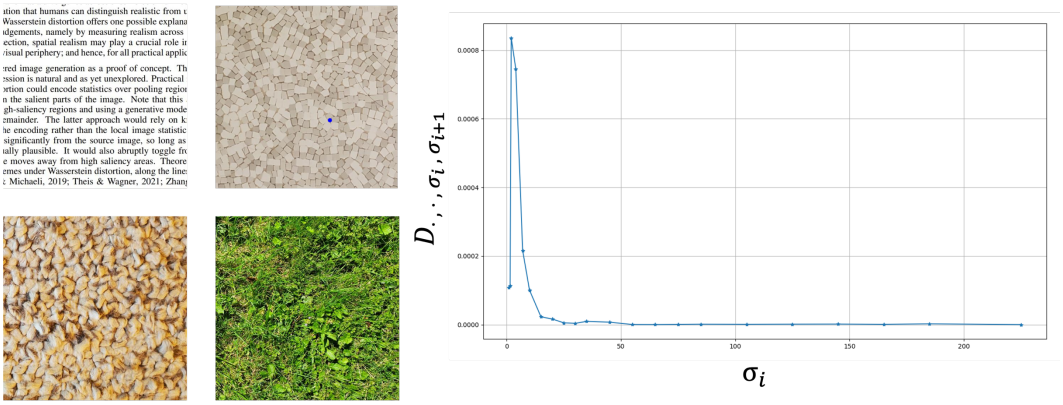

Figure 9: On the left are some examples of single-textural images, of which the intrinsic σ-map are mostly 0. On the right is an example of $\mathbf{d}$ of a particular pixel location in the second image (marked in blue), which reflects the behavior of $\mathbf{d}$ for most pixels that are assigned a σ value of 0. The problem happens because after the local perturbation stage, all empirical distributions with different σ's are nearly identical, wrongly categorizing the pixel location as non-textural.

Figure 10: The source image is on the left, with its intrinsic σ-map in the middle. We see that large σ values are assigned to the center of the text region, as the algorithm identifies the text area as textural. For the text to be perceptible, the entire text region should have small σ values. The reconstruction under the intrinsic σ-map is on the right, where the text area is illegible in the center.