Learning a Max-Entropy Diffusion Model for Visual Textures

Xinyuan Zhao
New York University
xz2556@nyu.edu

Eero P. Simoncelli
New York University & Flatiron Institute
eero.simoncelli@nyu.edu

Abstract

Visual textures, defined as spatially homogeneous image regions containing repeated elements (e.g. a field of grass or the bark of a tree), are prevalent in visual scenes and provide important cues for recognizing materials and objects. Existing texture models extract essential features from a single texture image, and can then generate high-quality samples that are visually similar to the original. However, their features are either hand-designed or based on a network pretrained for another purpose (e.g., object recognition). We develop a novel principled method for unsupervised learning of a set of statistics that are used to constrain a maximum entropy density model for the texture. We use training and sampling procedures derived from generative diffusion models. Our trained model is more compact (512 features), but generates texture images whose quality is as good as, and often better than, previous methods. We also demonstrate qualitative convexity of the representation space by generating samples that interpolate between two given textures.

Visual texture representation and synthesis methods have been extensively studied. Most existing texture models follow Julesz's seminal conjecture (Julesz, 1962): the appearance of a texture is determined by a set of local statistics that are measured by the human visual system. Two textures with the same statistics will have the same (or similar) appearance. These models (e.g., Gatys et al., 2015; Portilla and Simoncelli, 2000; Zhu et al., 1998) demonstrate their success by synthesizing images with statistics matching those of an original image, and showing that these are similar in appearance to the original. However, the statistics that they use are either hand-designed or based on a pretrained object-recognition network. Here, we develop a method for learning a relevant set of statistics, unsupervised, from a dataset of texture image, and sampling from the maximum-entropy density subject to these statistics. A detailed description of this work is available on arXiv.

Model

We aim to learn a parametric family of probability densities over $x \in \mathbb{R}^n$ (an n pixel image), where each density corresponds to a texture class, and individual texture images are samples. We assume a maximum entropy formulation (Jaynes, 1957), in which the density is defined by the following optimization, parameterized by $\mu \in \mathbb{R}^d$:

$$\max_p \mathbb{E}_p[-\log p(x)] \quad \text{s.t. } \mathbb{E}[f(x)] = \mu \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^d$ computes a set of d statistics to be learned from a dataset. The solution is:

$$p_\lambda(x) = \frac{1}{Z(\lambda)} \exp\left[-\lambda^T f(x)\right] \tag{2}$$

where $Z(\lambda)$ is a normalizing factor and $\lambda \in \mathbb{R}^d$ is chosen to satisfy the constraint $\mathbb{E}_{p_\lambda}[f(x)] = \mu$. The vectors λ and μ, both in $\mathbb{R}^d$, are uniquely determined by each other, and serve as a parameterization that controls which texture class the density represents.

Direct optimization of f and λ via maximum likelihood is intractable since $Z(\lambda)$ is an integral over the entire data space. Instead, we propose to learn f indirectly via denoising. Consider a noise-contaminated texture image $y = x + \sigma\varepsilon$, where ε is a sample of zero-mean identity-covariance Gaussian noise. A denoising function $\hat{\varepsilon}(y)$ trained to minimize the mean squared error in predicting ε will learn to compute (or approximate) the posterior mean, $\mathbb{E}(\varepsilon|y)$, which is proportional to the score (gradient of the log probability) of y (Miyasawa, 1961):

$$\mathbb{E}(\varepsilon|y) = -\sigma \nabla_y \log p(y) \tag{3}$$

We further assume that noisy y follows the distribution given by (2), with a different λ from the clean x. Then $\nabla_y \log p(y) = -\nabla_y[\lambda^T f(y)]$. Note that $Z(\lambda)$ disappears. We choose a log-normal distribution for σ, parameterize neural networks $f_\theta(y)$ and $\lambda_\phi(x, \sigma)$ and optimize:

$$\min_{\theta,\phi} \mathbb{E}_{x,\sigma,\varepsilon} \left\| \nabla_y[\lambda_\phi(x,\sigma)^T f_\theta(y)] - \varepsilon \right\|^2 \tag{4}$$

The multiplication by σ is absorbed into $\lambda_\phi(x, \sigma)$. For network architecture, we base $f_\theta(y)$ on UNet (Ronneberger et al., 2015) and $\lambda_\phi(x, \sigma)$ on ConvNeXt (Liu et al., 2022) with FiLM modulation (Perez et al., 2018). We set the number of statistics to $d = 512$.

This is a denoiser (or equivalently, ε-predictor) that operates at all noise levels σ, which provides a substrate for a generative diffusion model, the state-of-the-art image synthesis method (Ho et al., 2020; Ramesh et al., 2022; Rombach et al., 2022). Architecturally, we rely on

Extended abstract presented at the *9th Conference on Cognitive Computational Neuroscience*, New York, NY, USA, 2026.

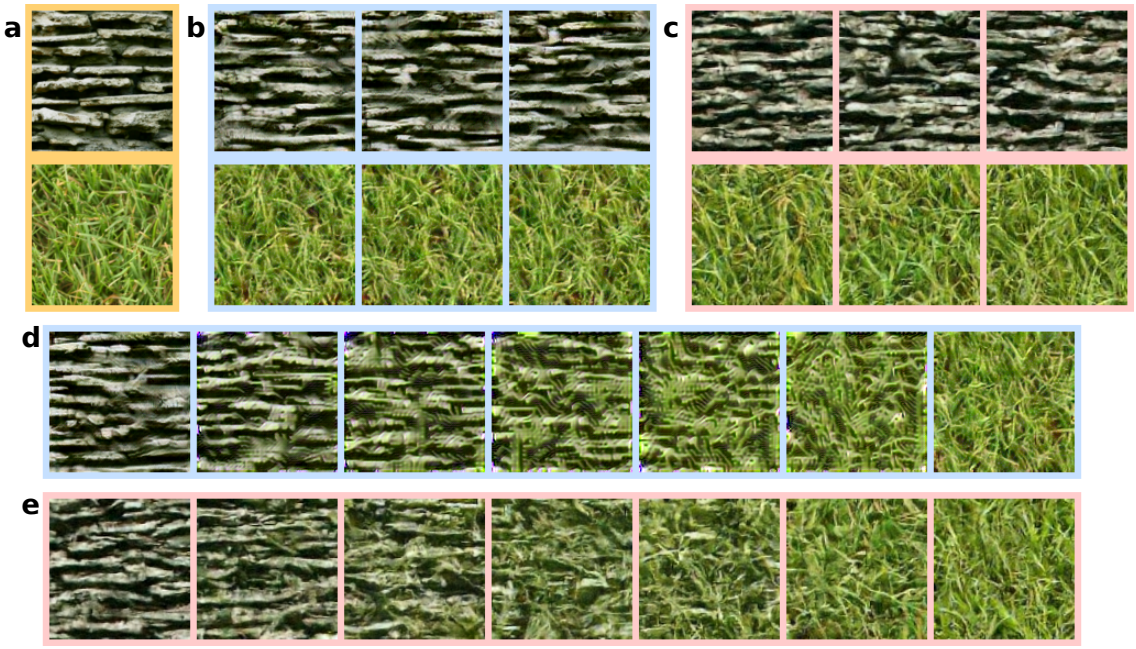

Figure 1: Texture resampling and interpolation. **(a)** Texture images that are not in the training set. **(b,c)** Resampling. **(d,e)** Interpolation. Panels (b,d) use statistics matching, and panels (c,e) use diffusion sampling.

two networks whose inputs are the same image up to additive noise, similar to some previous methods that use a diffusion model for unsupervised representation learning (Hudson et al., 2024; Mittal et al., 2023; Preechakul et al., 2022; Wang et al., 2023; Yang & Mandt, 2023).

For training, we generated a texture dataset by selecting one million 128x128 images patches from the ImageNet21K natural image dataset (Deng et al., 2009), using a hand-designed homogeneity criterion.

Results

We trained with random crops from our dataset for 5 epochs. We then use two methods for generating new texture images ("resampling") based on test image x_0:

- **Statistics matching**, solve $\min_x \|f(x) - f(x_0)\|^2$ (as in Gatys et al. (2015) and Portilla and Simoncelli (2000)).
- **Diffusion sampling**, use $\lambda = \lambda(x_0, \sigma)$ for each time step σ (Ho et al., 2020).

For high dimensional x such as images, the value of $f(x)$ concentrates around its mean μ (equivalence of ensembles), so the two methods should lead to similar results.

Additionally, we can empirically examine the convexity of these spaces by generating textures whose parameters lie between those of two existing textures x_a and x_b. This interpolation can be performed either μ or λ space:

$$\mu_i = \frac{N-i}{N} f(x_a) + \frac{i}{N} f(x_b), \qquad i = 0, \cdots, N \tag{5}$$

And similarly for λ. Then we use statistics matching and diffusion sampling to generate images from interpolated μ_i and λ_i, respectively. Results are shown in Figure 1.

We find that statistics matching has better resampling quality than diffusion matching, but the latter generates smoother interpolation. Subjective evaluation suggests our resampled textures are as good as, or better than, previous methods, while using a significantly smaller number of statistics (512 v.s. 176,640 in Gatys et al. (2015)). To evaluate quality and diversity, we compute the FID score (Heusel et al., 2017) on 9 test textures (Table 1). Except for one, the best-performing model is ours with either statistics matching or diffusion sampling.

Table 1: FID scores. From top to bottom: statistics matching for Gatys et al. (2015) model, statistics matching for our model, diffusion sampling for our model.

	Grass	Pebble	Star	Cloth	Rug	Flower	Marble	Rubber	Glitter
Gatys	175.03	**169.35**	209.10	126.79	273.30	74.03	135.17	135.52	96.19
Stat.	**102.58**	214.77	68.19	**41.70**	105.96	80.49	137.79	136.71	**20.82**
Diff.	221.30	192.94	**48.78**	66.25	**55.49**	**67.14**	**46.68**	**47.24**	36.82

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Gatys, L., Ecker, A. S., & Bethge, M. (2015). Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, *28*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, *30*.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.

Hudson, D. A., Zoran, D., Malinowski, M., Lampinen, A. K., Jaegle, A., McClelland, J. L., Matthey, L., Hill, F., & Lerchner, A. (2024). Soda: Bottleneck diffusion models for representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23115–23127.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, *106*(4), 620.

Julesz, B. (1962). Visual pattern discrimination. *IRE transactions on Information Theory*, *8*(2), 84–92.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Mittal, S., Abstreiter, K., Bauer, S., Schölkopf, B., & Mehrjou, A. (2023). Diffusion based representation learning. *International conference on machine learning*, 24963–24982.

Miyasawa, K. (1961). An empirical bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist*, *38*(181-188).

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). Film: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI conference on artificial intelligence*, *32*(1).

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, *40*, 49–70.

Preechakul, K., Chatthee, N., Wizadwongsa, S., & Suwajanakorn, S. (2022). Diffusion autoencoders: Toward a meaningful and decodable representation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10619–10629.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, 234–241.

Wang, Y., Schiff, Y., Gokaslan, A., Pan, W., Wang, F., De Sa, C., & Kuleshov, V. (2023). Infodiffusion: Representation learning using information maximizing diffusion models. *International Conference on Machine Learning*, 36336–36354.

Yang, R., & Mandt, S. (2023). Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems*, *36*, 64971–64995.

Zhu, S. C., Wu, Y., & Mumford, D. (1998). Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, *27*(2), 107–126.