Generative models for when inference gets complicated.

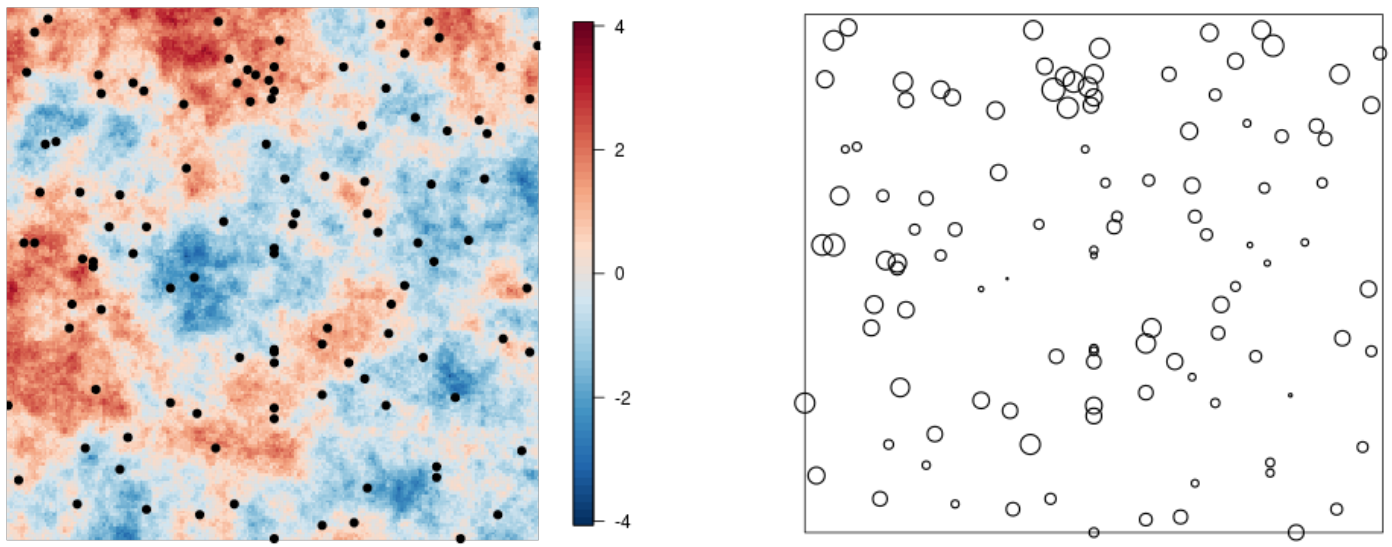
Gabriel V. Cardoso and Mike Pereira (Ecole des Mines - PSL)

April 14, 2025

The problem

Motivation

- Data: Observations $\{x(s_i)\}_{i=1}^n$ of a variable x at n locations $s_1, \ldots, s_n$ of a domain $\mathcal{D} \subset \mathbb{R}^d$



Left: Variable z. Right : Observations

$\rightarrow$ **Goal: Predict x at a new location $s_0 \in \mathcal{D}$**

Motivation

Geostatistical paradigm: over the spatial domain $\mathcal{D}$

Gaussian Random Field		Observed variable
$X : \{X(s) : s \in \mathcal{D}\}$	$\xrightarrow{\text{Realization}}$	$x : \{x(s) : s \in \mathcal{D}\}$
High correlation		High *similarity*

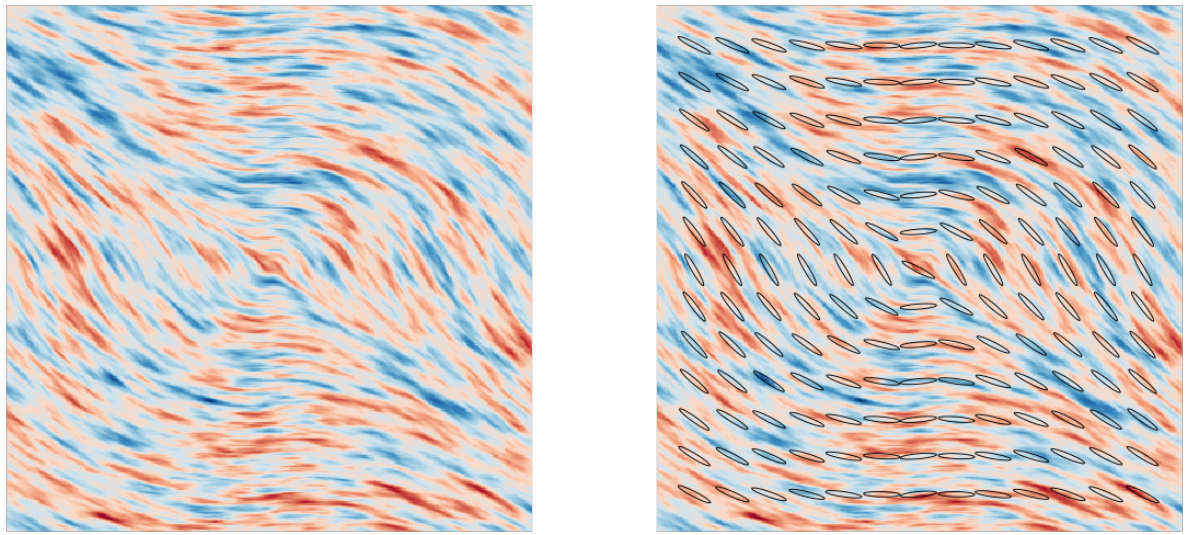
- Finite-dimensional distributions fully specified by mean function + covariance function $\rightarrow$ Parametrized by θ
- Analytic formulas for predictive posterior distribution

$$(X(s_0)|X(s_1), \ldots, X(s_n), \theta)$$
$$\sim \mathcal{N}(\text{Kriging prediction at } s_0, \text{Kriging variance at } s_0)$$

$\rightarrow$ BUT Need to **estimate parameters** θ from data to compute it!

Motivation

- Parameter estimation relatively straightforward for istropic models (Max-likelihood, MCMC,...)
- But what if we consider non-stationary Gaussian random fields (with locally varying parameters) ?



- Maps of parameters $\Rightarrow$ Complex parameter space $\Rightarrow$ Max-Likelihood/MCMC harder to converge
- INLA approach (i.e. Laplace approximation) implemented for less complex models (varying range, variance)

Motivation

- Idea : Use generative models to perform Bayesian predictions
- Our setting :
  - Set prior distribution on parameters θ
  - Use generative models to learn the corresponding prior distribution of GRFs
  - Use conditioning of generative model to sample from the posterior predictive distribution $(X(s_0)|X(s_1), \ldots, X(s_n))$ ($\rightarrow$ Marginalized over θ !)

Fast introduction to score-based generative models

The original miracle!

Figure: $X_0 \sim \pi$.

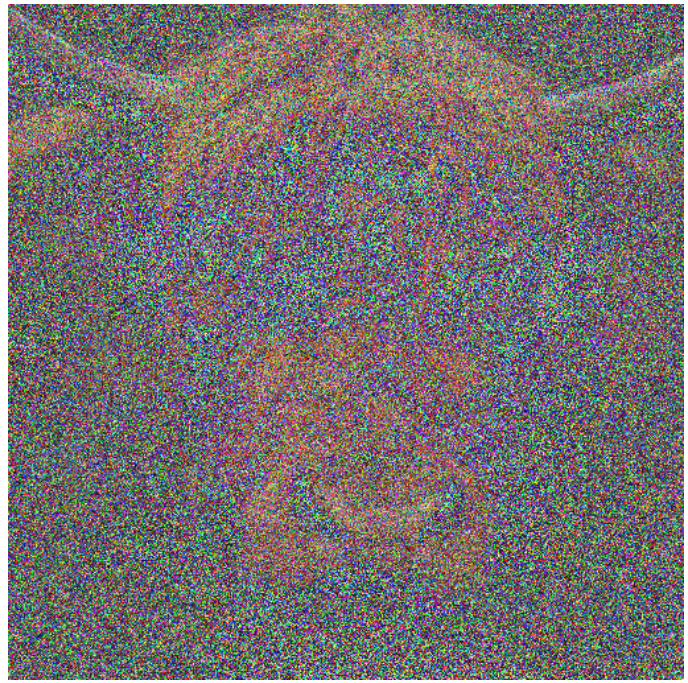


Figure: $X_\sigma = X_0 + \sigma Z$. [1]1

[1] $Z \sim \mathcal{N}(0, \mathrm{I}), X_0 \perp Z$.

The original miracle!

Figure: $X_0 \sim \pi$.

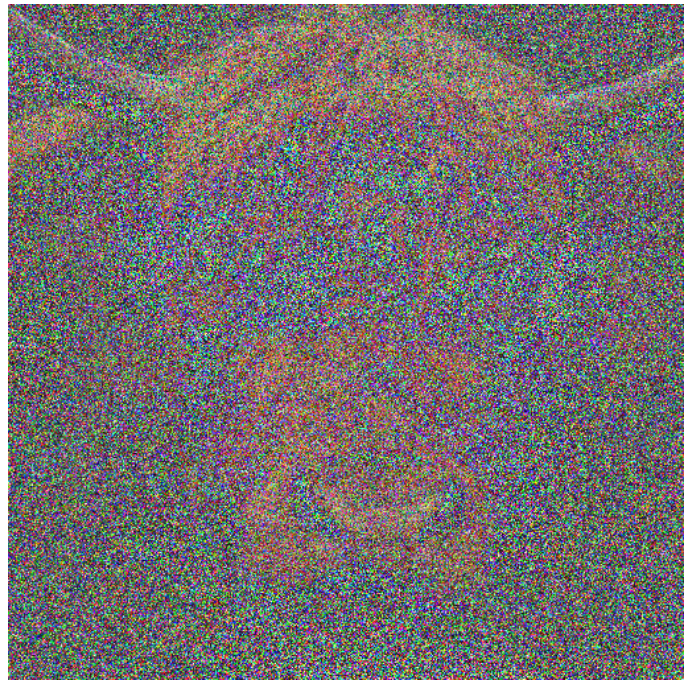


Figure: $X_\sigma = X_0 + \sigma Z$. [1]1

- D_{θ_0} untrained CNN,
- $L_\sigma(\theta) = \|D_\theta(X_\sigma) - X_\sigma\|^2$, $L_0(\theta) = \|D_\theta(X_\sigma) - X_0\|^2$,
- $\theta_t = \theta_{t-1} - \nabla L_\sigma(\theta_{t-1})$

[1] $Z \sim \mathcal{N}(0, \mathrm{I})$, $X_0 \perp Z$.

The original miracle!

Figure: $X_0 \sim \pi$.

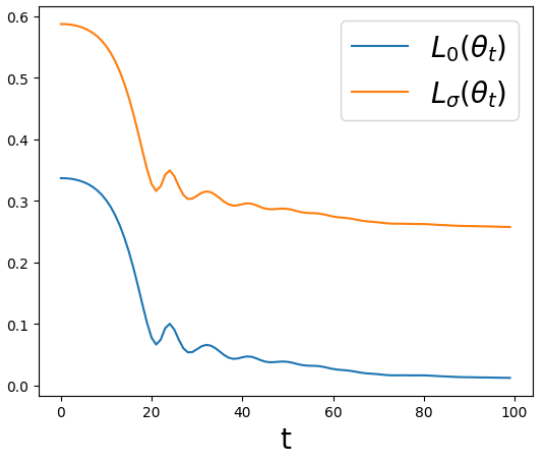
Figure: $X_\sigma = X_0 + \sigma Z$. [1]1

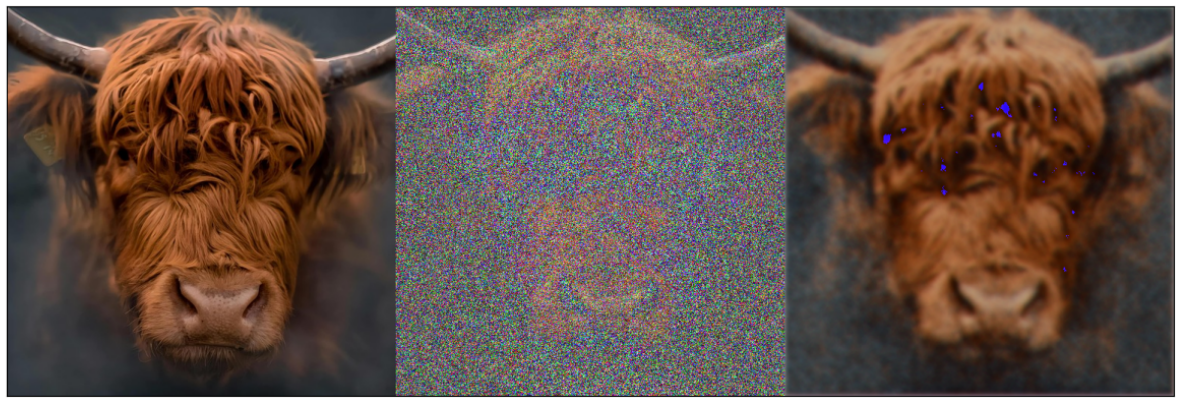
- D_{θ_0} untrained CNN,
- $L_\sigma(\theta) = \|D_\theta(X_\sigma) - X_\sigma\|^2$, $L_0(\theta) = \|D_\theta(X_\sigma) - X_0\|^2$,
- $\theta_t = \theta_{t-1} - \nabla L_\sigma(\theta_{t-1})$

What happens to $\{L_\sigma(\theta_t)\}_{t=1}^N$ and $\{L_0(\theta_t)\}_{t=1}^N$?

[1] $Z \sim \mathcal{N}(0, \mathrm{I})$, $X_0 \perp Z$.

The original miracle [2].





[2]This experiment is inspired by Ulyanov et al., "Deep Image Prior", 2018

Noising and Denoising

Figure: $X_0 \sim \pi$.

Figure: $X_t = X_0 + \sigma_t Z_t$. [3]1

Suppose we have an increasing noise schedule

$$\sigma : [0, T] \to \sigma_t \in \mathbb{R}_+ \, .$$

Can we reverse the procedure?

[1] $Z_t \perp X_0, Z_t \sim \mathcal{N}(0, \mathrm{I})$.

Noising and Denoising

Suppose we have an increasing noise schedule

$$\sigma : [0, T] \to \sigma_t \in \mathbb{R}_+ \, .$$

1. Noising kernel:

$$p_{t|0}(x_t|x_0) = \mathcal{N}(x_t; x_0, \sigma_t^2 \, \mathrm{I}) \, ,$$

2. Marginal:

$$p_t(x_t) = \int p_{t|0}(x_t|x_0)\pi(\mathrm{d}x_0) \, .$$

Our goal is to go backwards:

$$p_T \to p_{T-\Delta t} \to \cdots \to p_0 \, .$$

Figure: $X_t = X_0 + \sigma_t Z_t$. [4]1

Generative models by score matching[5]

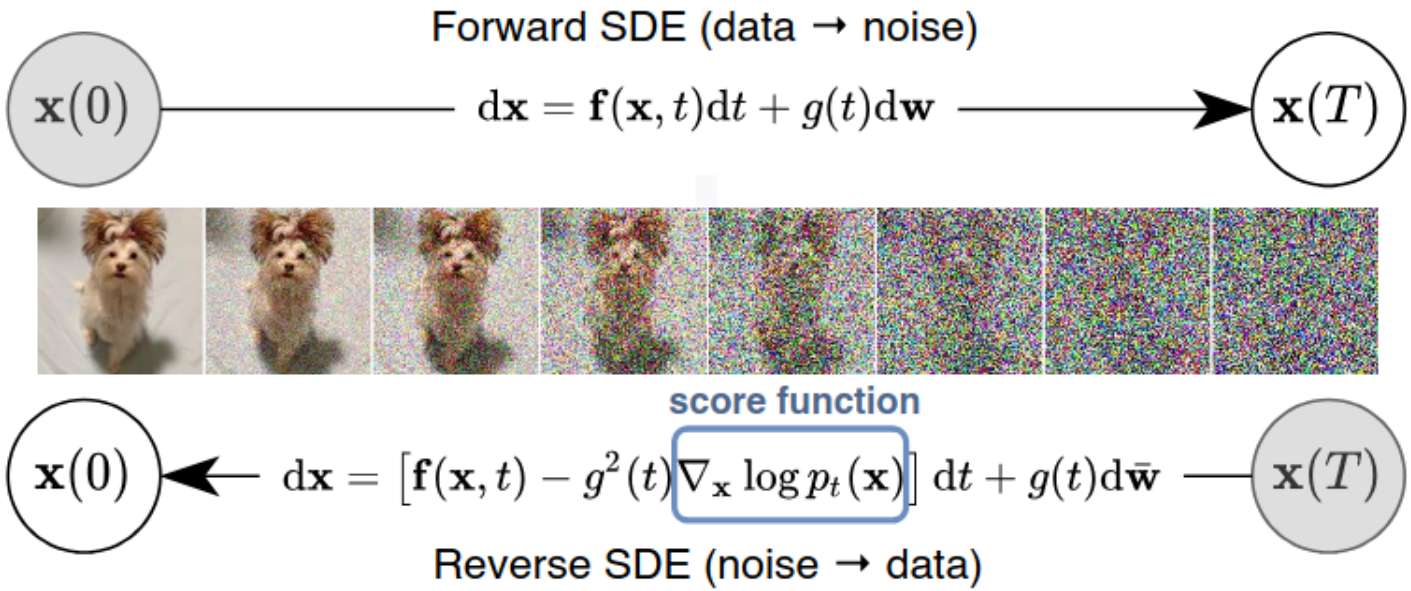


Figure: Illustration of noising and denoising with SDEs.

[4]From Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations", 2021.

[5]Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations", 2021.

Generative models by score matching[5]

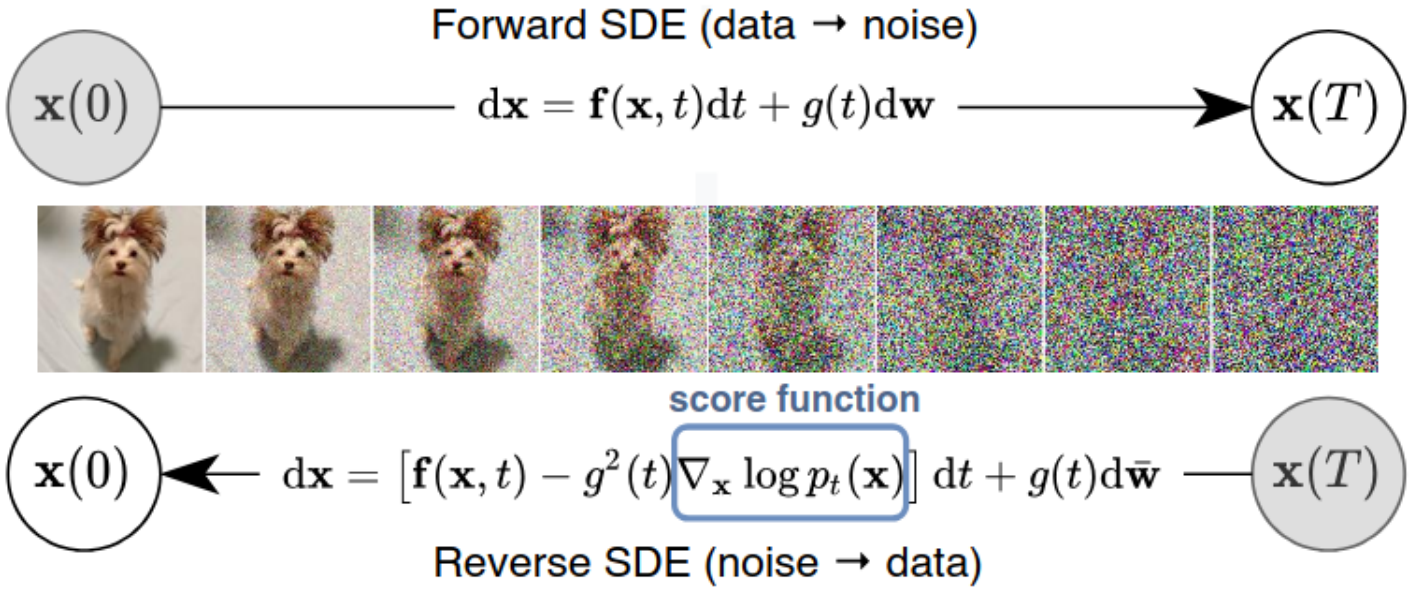


Figure: Illustration of noising and denoising with SDEs.

Our p_t is the time t marginal of the SDE

$$\mathrm{d}x = \mathrm{d}w\,.$$

It admits the backward decomposition

$$\mathrm{d}x = -\nabla \log p_t(x)\mathrm{d}_t + \mathrm{d}\bar{w}\,.$$

[4]From Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations", 2021.

[5]Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations", 2021.

Theoretical guarantees

We can control

- $\mathsf{KL}(\mathsf{q}_{0:T}(x_{0:T}) \parallel \mathsf{p}_{0:T}(x_{0:T}))$[6]
- $\mathrm{W}_2(\mathsf{p}_{0:T}(x_{0:T}), \mathsf{q}_{0:T}(x_{0:T}))$.[7]

[6] Conforti et al., *KL Convergence Guarantees for Score Diffusion Models under Minimal Data Assumptions*, 2024.

[7] Gao et al., *Wasserstein Convergence Guarantees for a General Class of Score-Based Generative Models*, 2025.

Denoising: Optimal denoiser and the score

Consider the mean squared error (MSE)

$$\mathrm{MSE}\left(f, \sigma_t^2\right) := \mathbb{E}_{\mathsf{q}_{t,0}}\left[\|f(X_t) - X_0\|^2\right].$$

[8] Vincent, "A Connection Between Score Matching and Denoising Autoencoders", 2011.

Denoising: Optimal denoiser and the score

Consider the mean squared error (MSE)

$$\mathrm{MSE}\left(f, \sigma_t^2\right) := \mathbb{E}_{\mathsf{q}_{t,0}}\left[\|f(X_t) - X_0\|^2\right].$$

The optimal denoiser w.r.t the MSE is,

$$x_t \to \mathbb{E}\left[X_0 \,|\, X_t = x_t\right] \in \operatorname*{argmin}_{f \in L^2(\mathsf{q}_t)} \mathrm{MSE}\left(f, \sigma_t^2\right).$$

[8] Vincent, "A Connection Between Score Matching and Denoising Autoencoders", 2011.

Denoising: Optimal denoiser and the score

Consider the mean squared error (MSE)

$$\mathrm{MSE}\left(f, \sigma_t^2\right) := \mathbb{E}_{\mathsf{q}_{t,0}}\left[\|f(X_t) - X_0\|^2\right].$$

The optimal denoiser w.r.t the MSE is,

$$x_t \to \mathbb{E}\left[X_0 \,|\, X_t = x_t\right] \in \underset{f \in L^2(\mathsf{q}_t)}{\operatorname{argmin}} \mathrm{MSE}\left(f, \sigma_t^2\right).$$

It is possible to show that[8]

$$\mathbb{E}\left[X_0 \,|\, X_t = x_t\right] = x_t + \sigma_t^2 \underbrace{\nabla \log \mathsf{q}_t(x_t)}_{\text{score}}.$$

[8] Vincent, "A Connection Between Score Matching and Denoising Autoencoders", 2011.

Denoising: Optimal denoiser and the score

Consider the mean squared error (MSE)

$$\mathrm{MSE}\left(f, \sigma_t^2\right) := \mathbb{E}_{\mathsf{q}_{t,0}}\left[\|f(X_t) - X_0\|^2\right].$$

The optimal denoiser w.r.t the MSE is,

$$x_t \to \mathbb{E}\left[X_0 \,|\, X_t = x_t\right] \in \underset{f \in L^2(\mathsf{q}_t)}{\operatorname{argmin}} \mathrm{MSE}\left(f, \sigma_t^2\right).$$

It is possible to show that[8]

$$\mathbb{E}\left[X_0 \,|\, X_t = x_t\right] = x_t + \sigma_t^2 \underbrace{\nabla \log \mathsf{q}_t(x_t)}_{\text{score}}.$$

Learn a denoiser $\Longleftrightarrow$ Learn the score .

[8] Vincent, "A Connection Between Score Matching and Denoising Autoencoders", 2011.

Backward sampling from Ho et al., "Denoising diffusion probabilistic models", 2020

Algorithm Backward sampling

1: **procedure** SAMPLING
Input: N.
Output: X_0.
2: $X_N \sim \mathcal{N}(0, \sigma_T^2 \mathrm{I})$.
3: **for** $i = N-1, \cdots, 0$ **do**
4: $t_i = (i/N)T$
5: $Z_i \sim \mathcal{N}(0, \mathrm{I})$.
6: $X_i = X_{i+1} + \frac{\sigma_{t_i}^2 - \sigma_{t_{i+1}}^2}{\sigma_{t_{i+1}}^2}(X_{i+1} - D_\theta(X_{i+1}, \sigma_{t_{i+1}})) + \sqrt{\frac{\sigma_{t_i}^2}{\sigma_{t_{i+1}}^2}\left(\sigma_{t_{i+1}}^2 - \sigma_{t_i}^2\right)} Z_i$.
7: **end for**
8: **Return:** X_0
9: **end procedure**

DDPM: Illustration

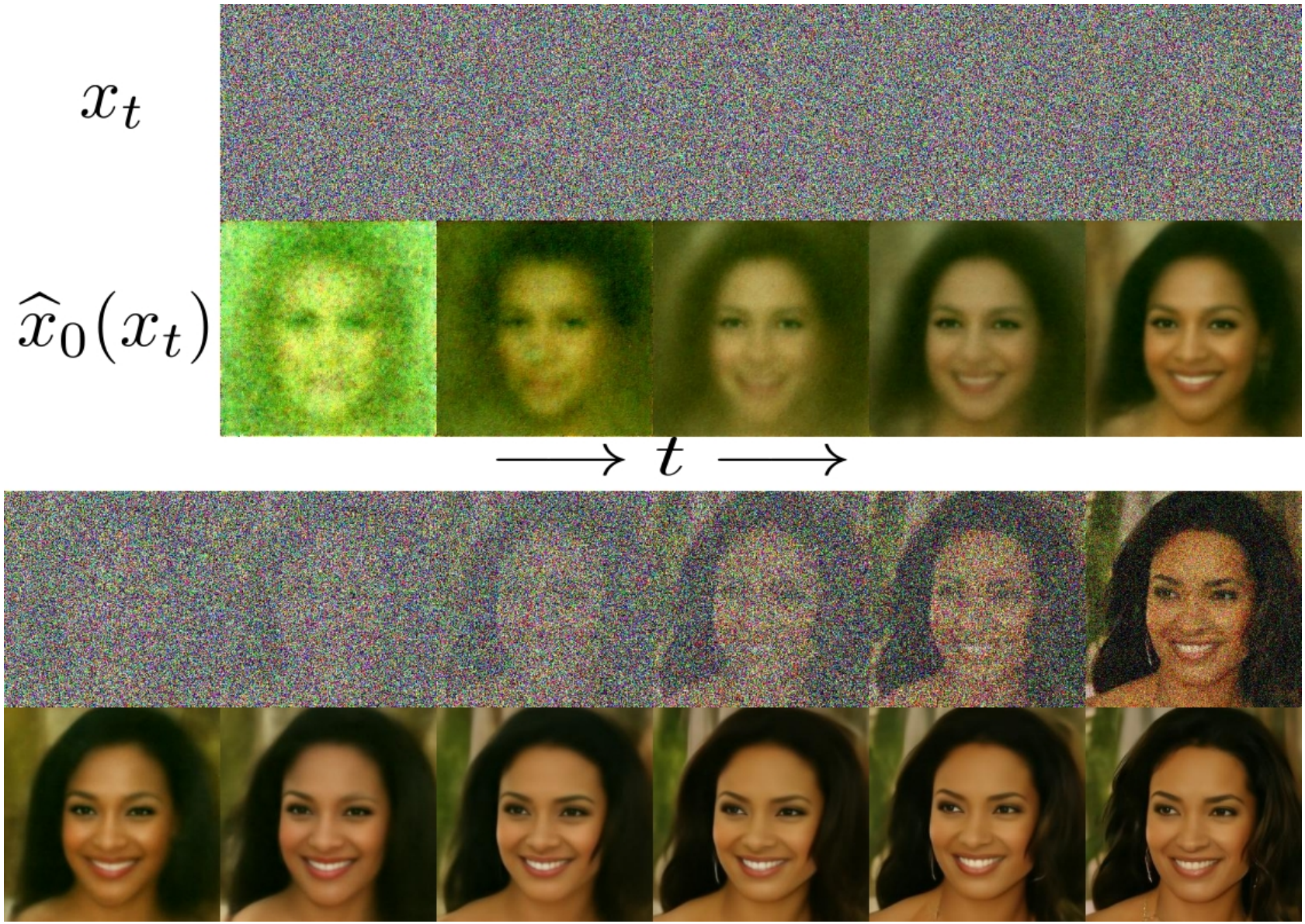


Figure: Illustration of DDPM on CelebA dataset.

Denoising with UNETs

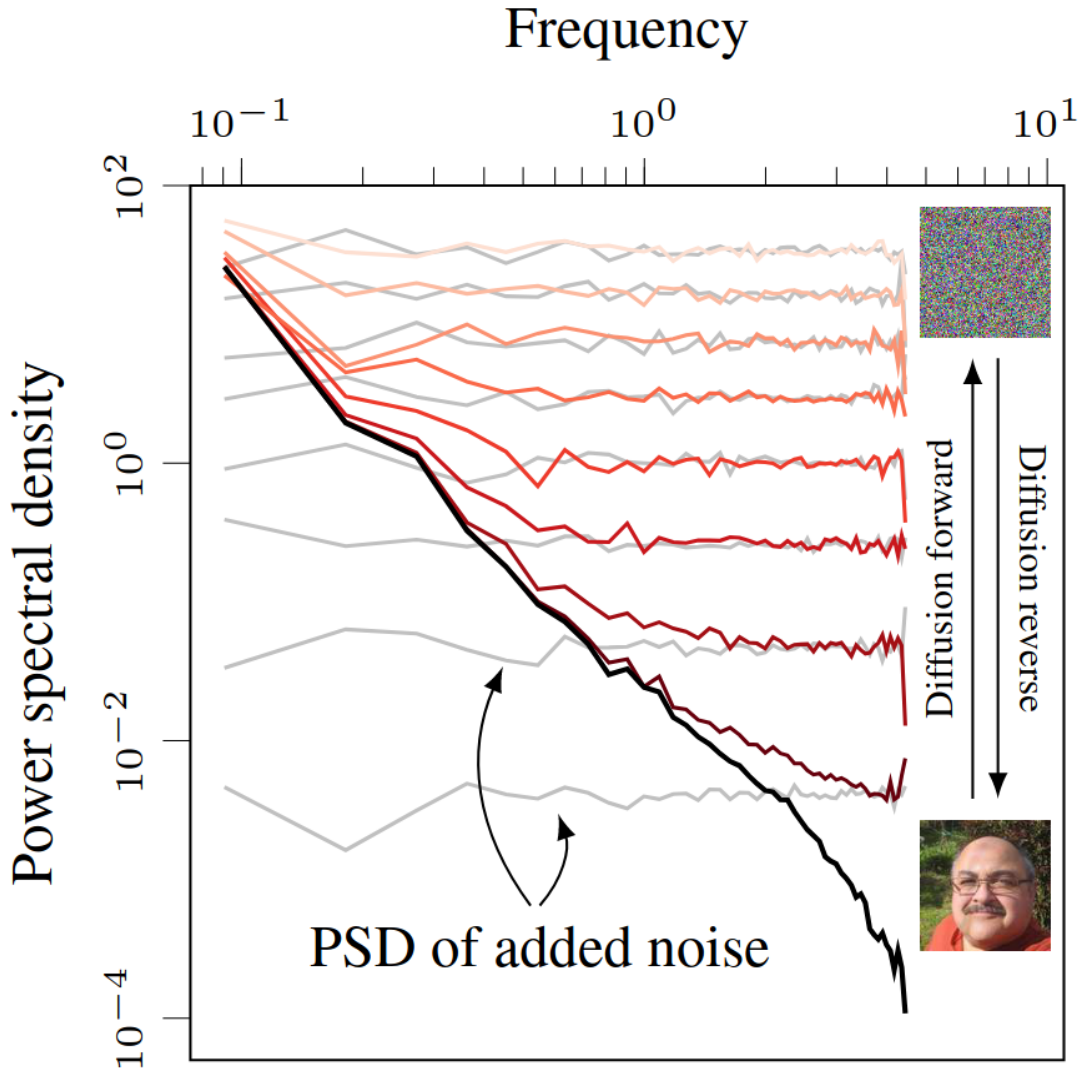


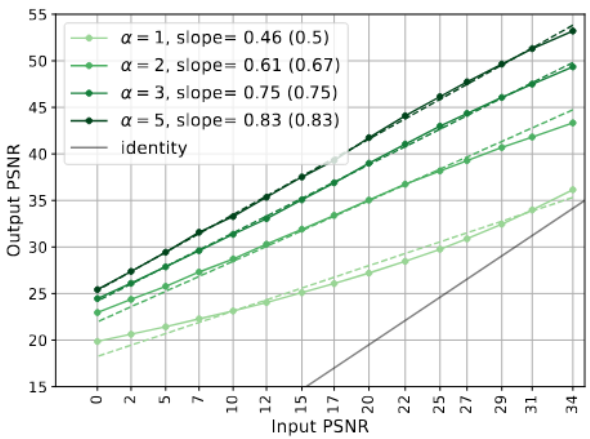
Figure: Illustration from Severi Rissanen et al. "Generative Modelling with Inverse Heat Dissipation". In: *The Eleventh International Conference on Learning Representations*. Sept. 2022. (Visited on 11/13/2024).

Denoising with UNETs: Geometry-adapted harmonic representation

Linear denoising in an adaptative basis:

$$\widehat{x}_0\left(x_t;\theta\right) = \underset{\lambda,\mathcal{B}\in\mathcal{D}}{\operatorname{argmin}} \sum_{e_k\in\mathcal{B}} \lambda_k(x_t) e_k(x_t)$$

Denoising with Unets in C^α images



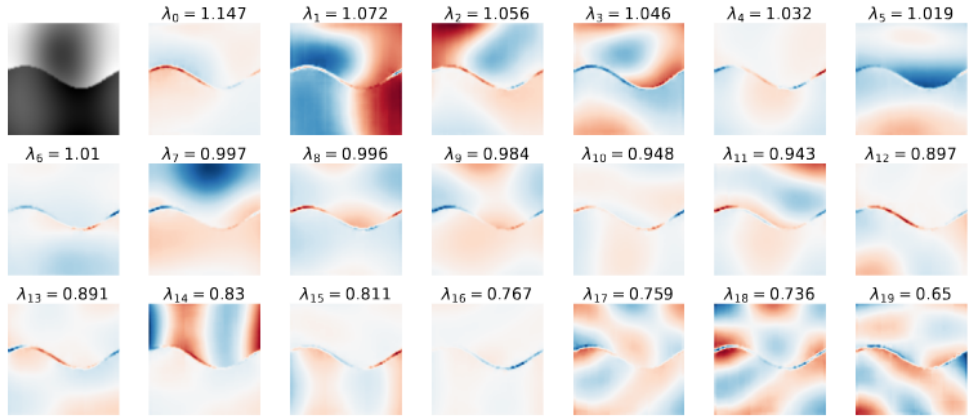


Figure: Illustration from Zahra Kadkhodaie et al. "Generalization in diffusion models arises from geometry-adaptive harmonic representation". en. In: *The Twelfth International Conference on Learning Representations*. 2023.

$$\mathrm{MSE}\left(f, \sigma\right) \propto \sigma^{2\frac{\alpha}{\alpha+1}}$$

Denoising with Unets: Celeb

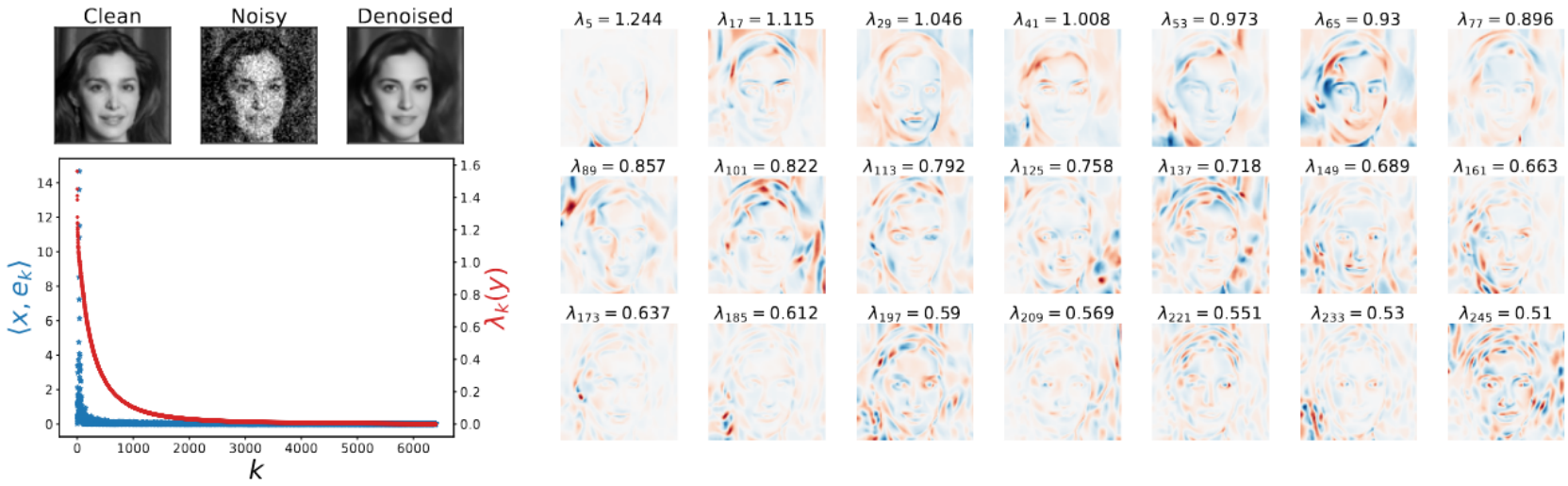


Figure: Illustration from Zahra Kadkhodaie et al. "Generalization in diffusion models arises from geometry-adaptive harmonic representation". en. In: *The Twelfth International Conference on Learning Representations*. 2023.

Denoising with UNETs

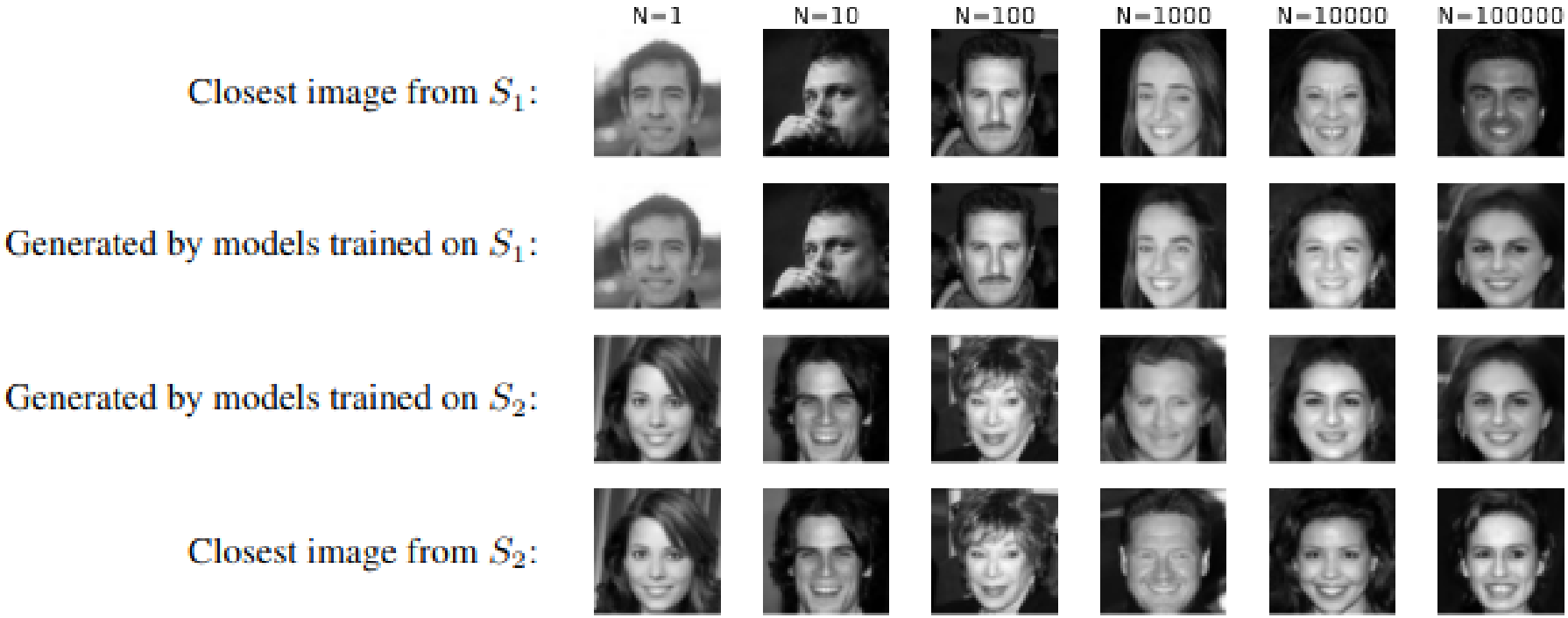


Figure: Illustration from Zahra Kadkhodaie et al. “Generalization in diffusion models arises from geometry-adaptive harmonic representation”. en. In: *The Twelfth International Conference on Learning Representations*. 2023.

Struggles and Results

Quizz

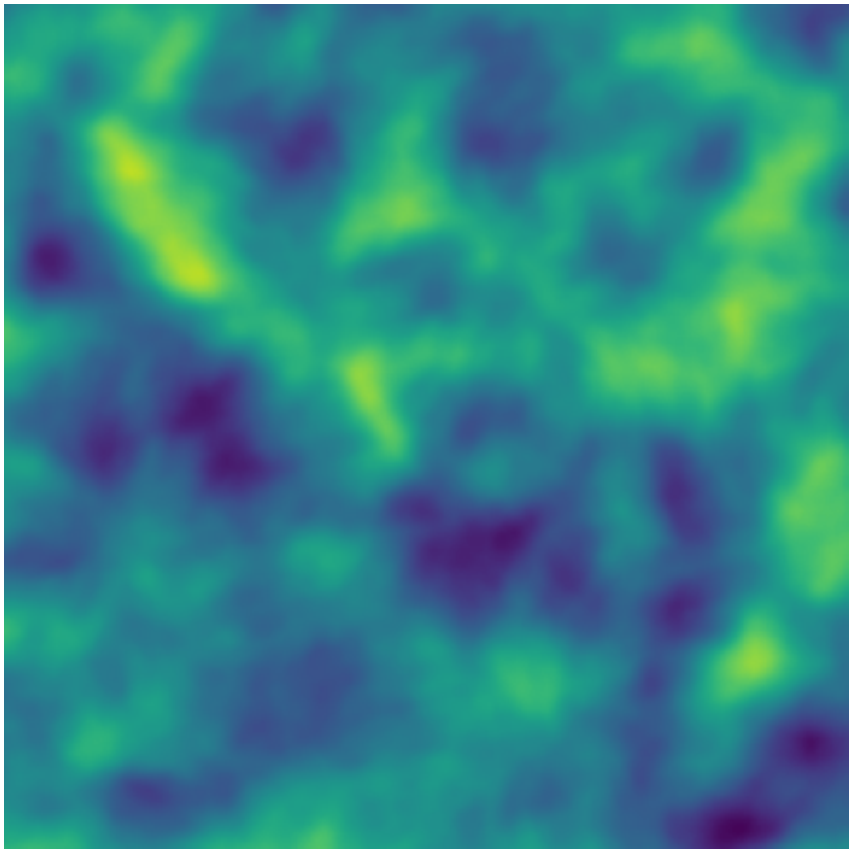


Figure: 0

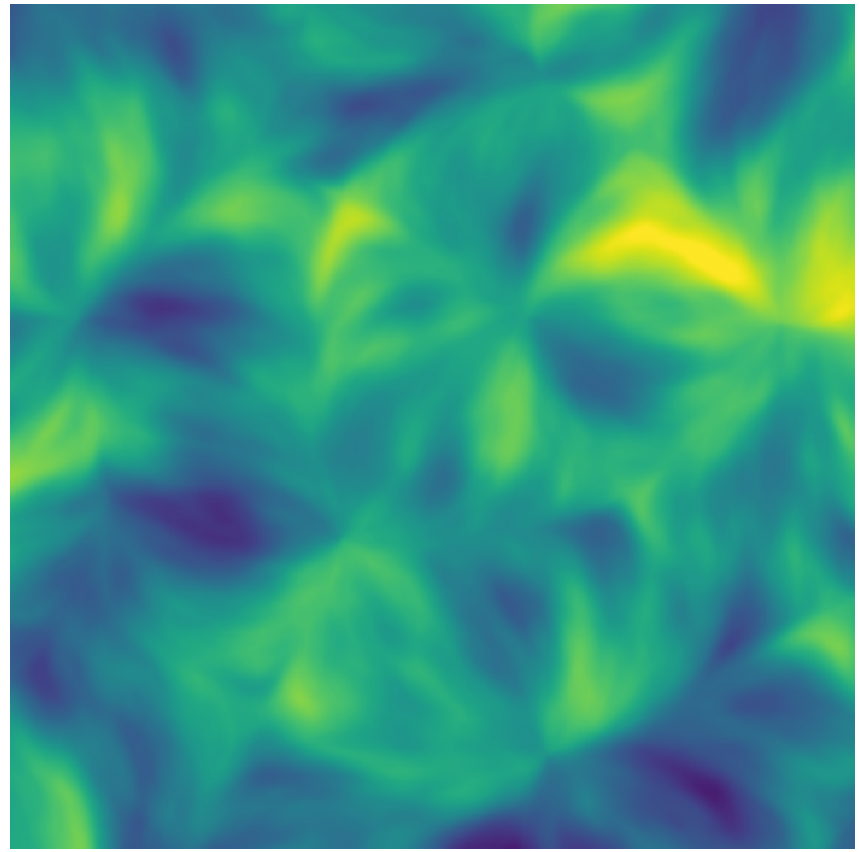


Figure: 1

Quizz

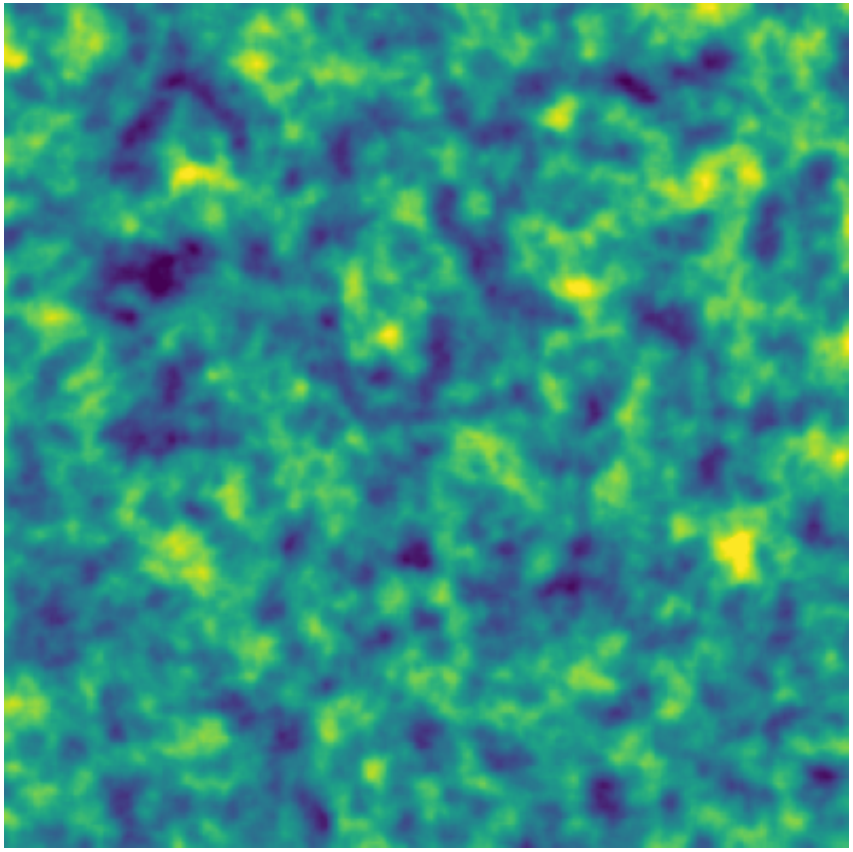


Figure: 0

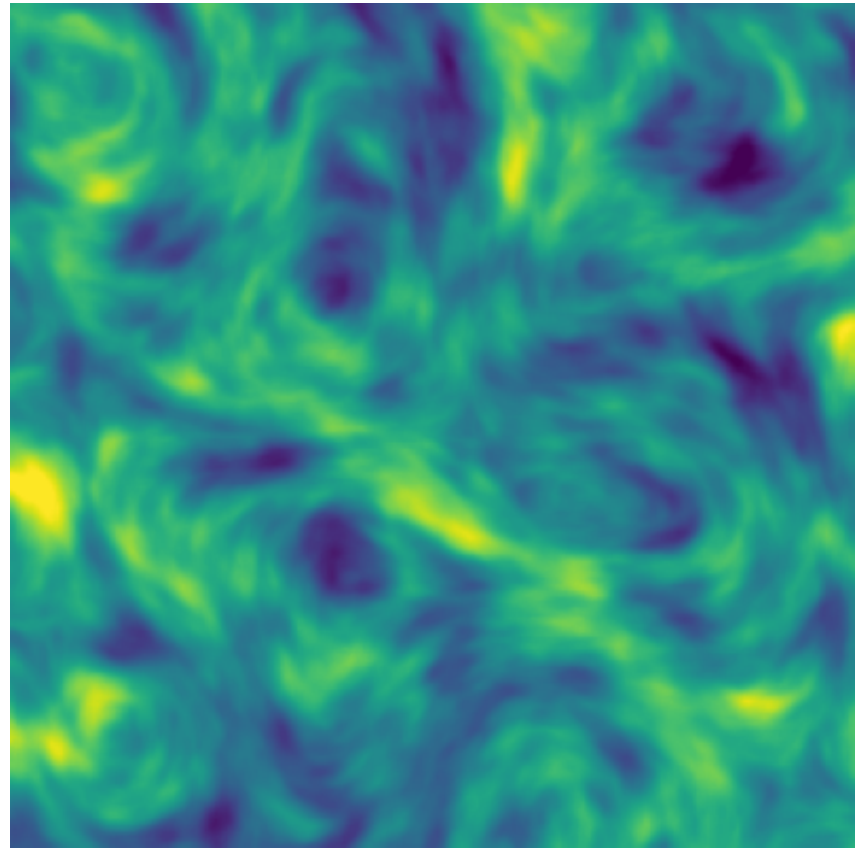


Figure: 1

Quizz

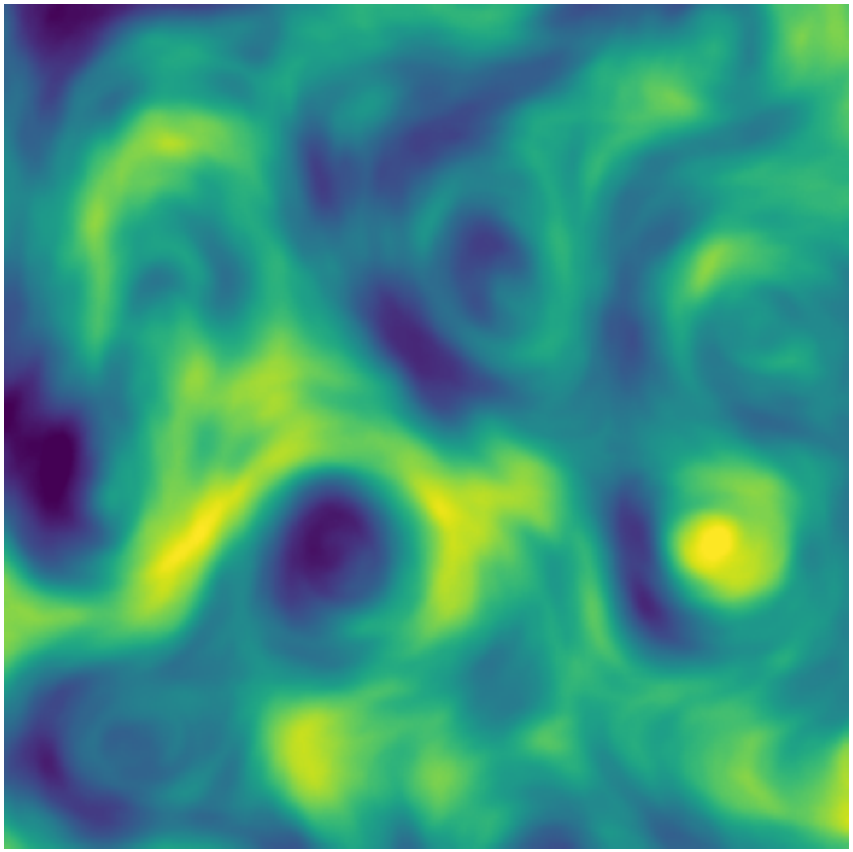


Figure: 0

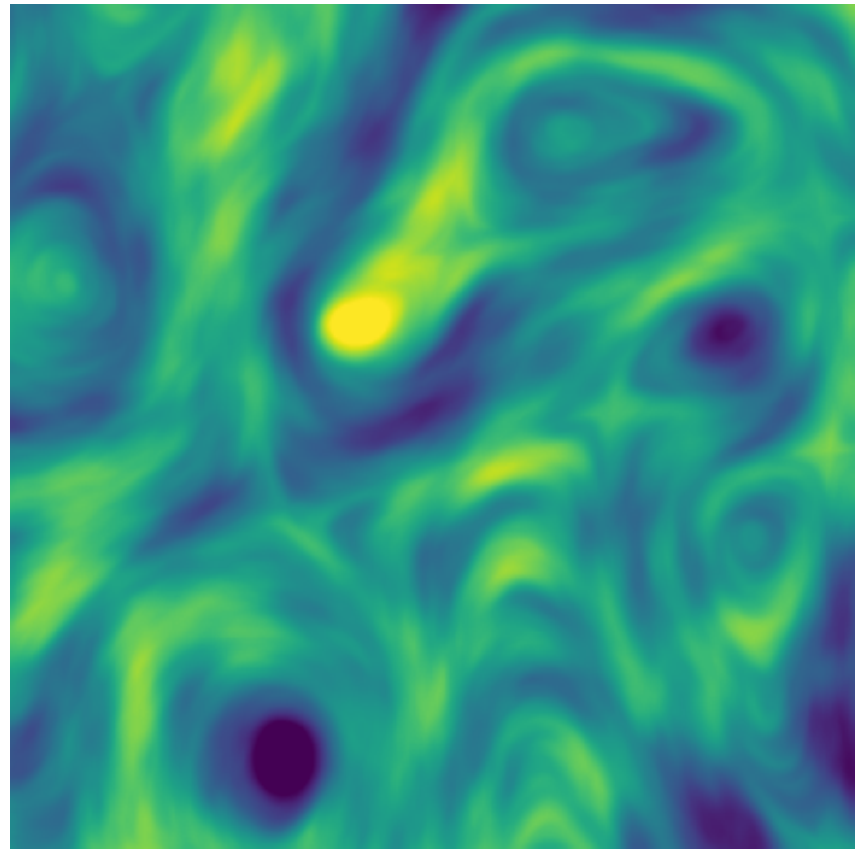


Figure: 1

Quizz

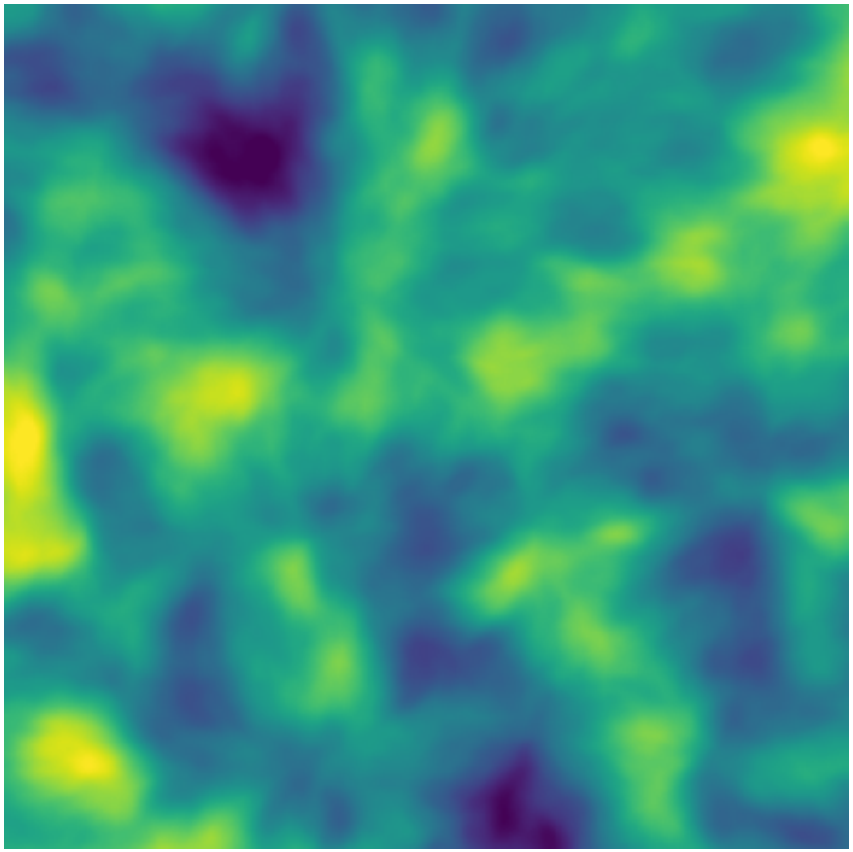


Figure: 0

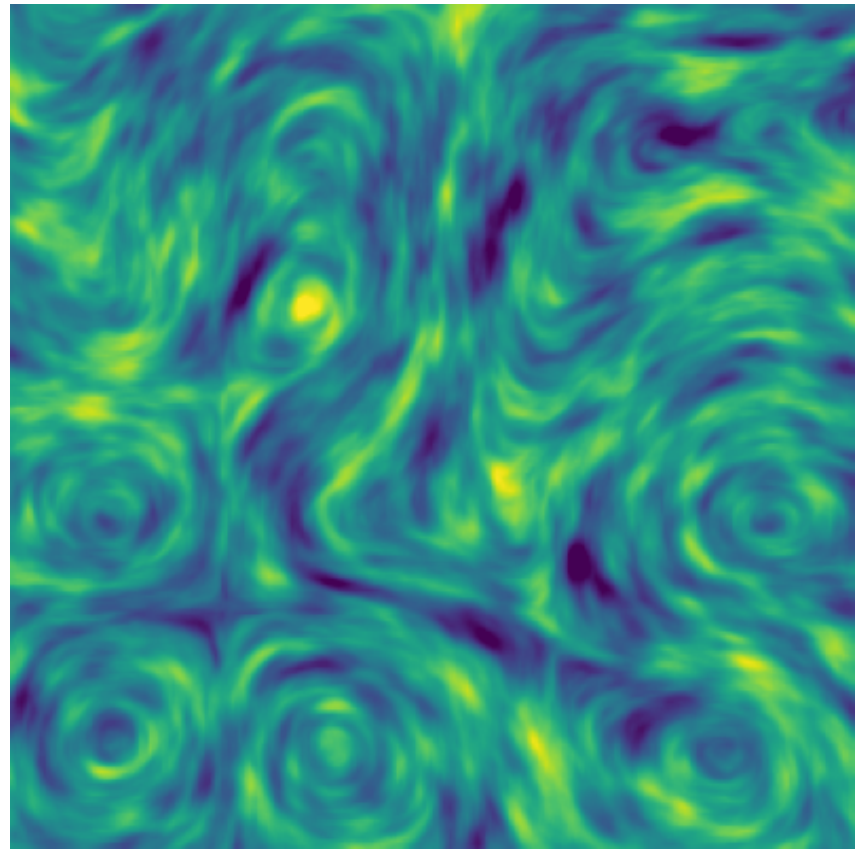


Figure: 1

Quizz

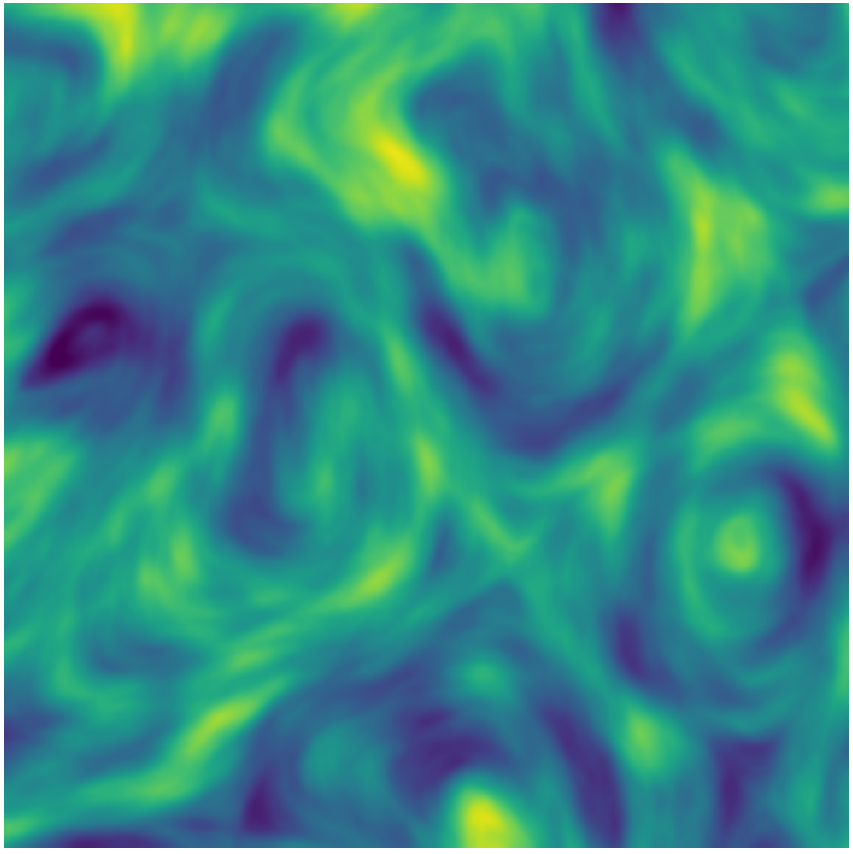


Figure: 0

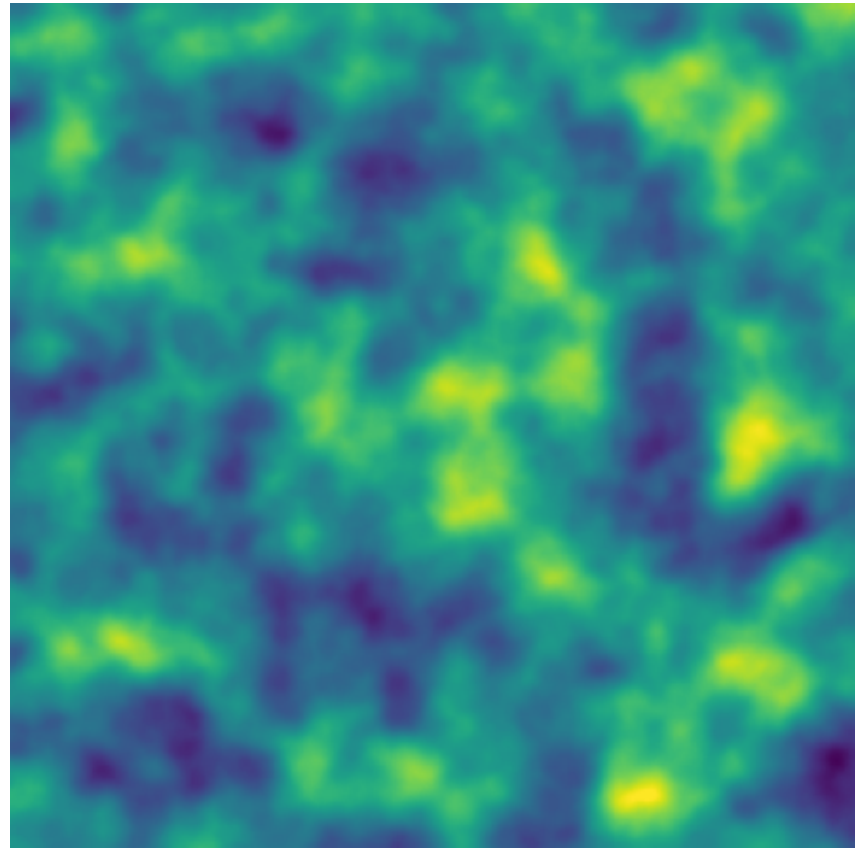


Figure: 1

Quizz

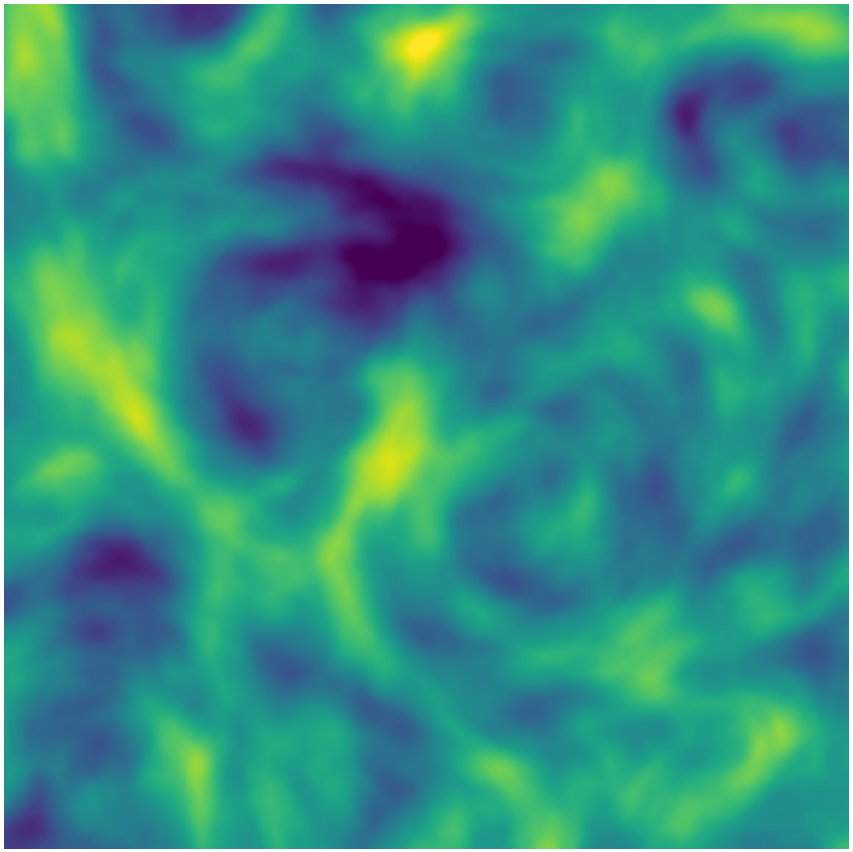


Figure: 0

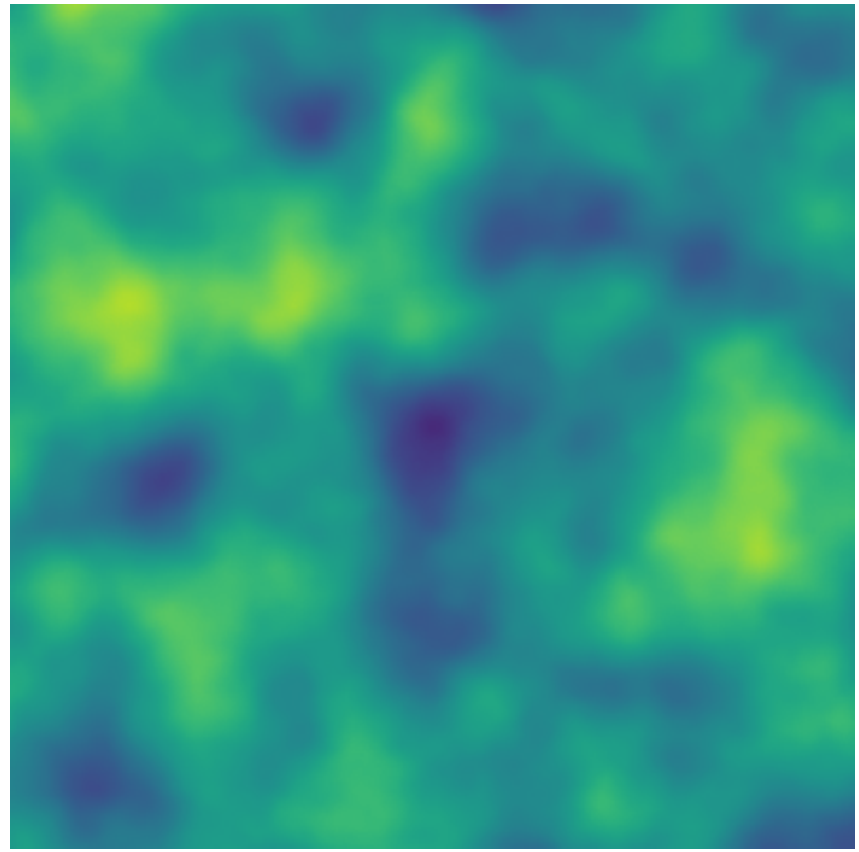
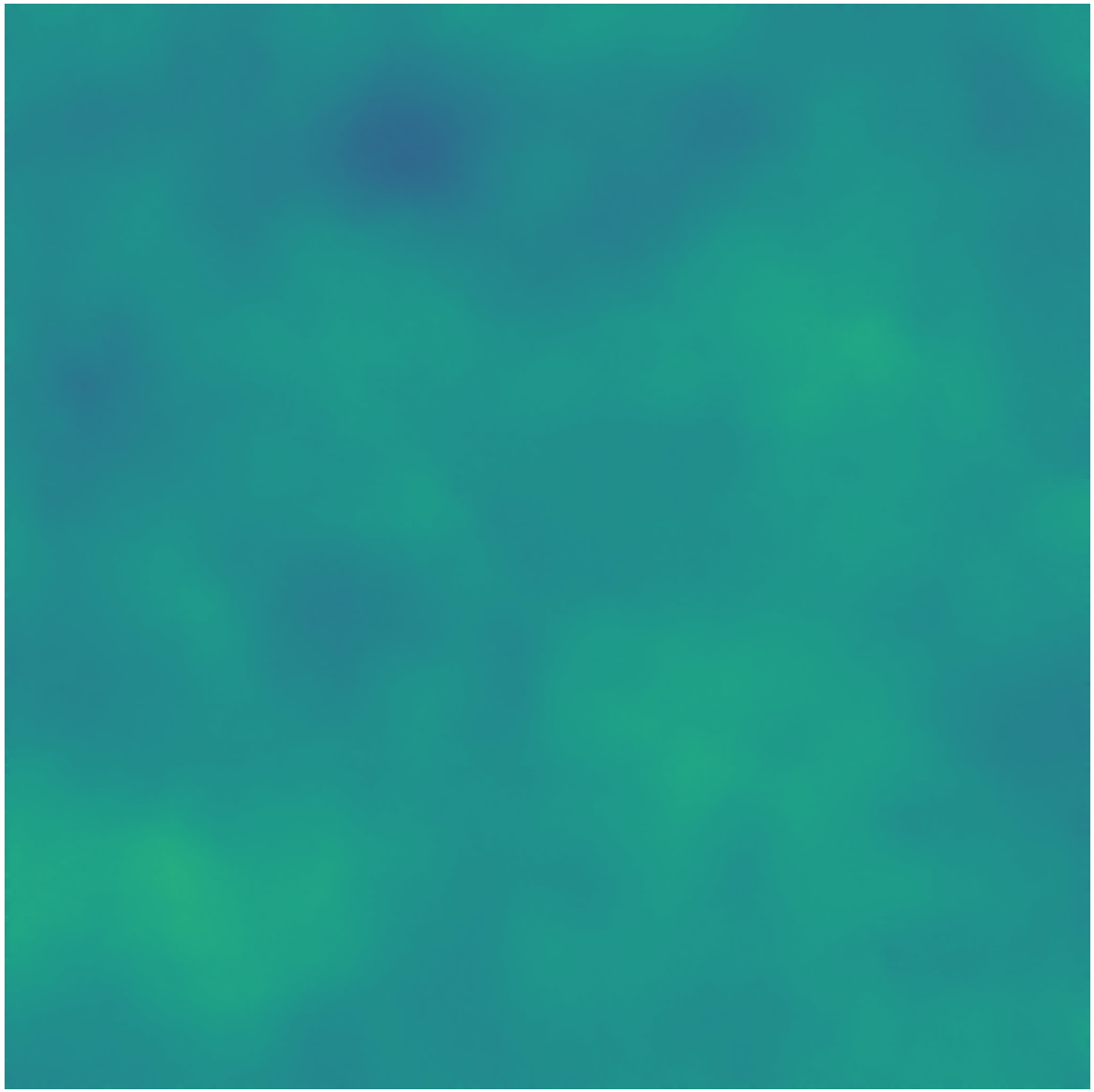


Figure: 1

Illustration



How to judge? Pt.1 Max-Sliced Wasserstein

Wasserstein distances

Let $\mu, \nu \in \mathcal{P}_p$. We define

$$\mathrm{W}_p(\mu, \nu)^p := \inf_{\gamma \in C(\mu,\nu)} \int \|x - y\|^p \gamma(\mathrm{d}x, \mathrm{d}y)\,, \tag{1}$$

[9] See Nietert et al., "Statistical, Robustness, and Computational Guarantees for Sliced Wasserstein Distances", 2022 for guarantees.

How to judge? Pt.1 Max-Sliced Wasserstein

Wasserstein distances

Let $\mu, \nu \in \mathcal{P}_p$. We define

$$\mathrm{W}_p(\mu,\nu)^p := \inf_{\gamma \in C(\mu,\nu)} \int \|x-y\|^p \gamma(\mathrm{d}x, \mathrm{d}y)\,, \tag{1}$$

$$\overline{\mathrm{W}}_p(\mu,\nu) = \max_{\theta \in \mathbb{S}^{d-1}} \mathrm{W}_p(\mathfrak{p}^\theta \# \mu, \mathfrak{p}^\theta \# \nu)\,. \tag{2}$$

[9] See Nietert et al., "Statistical, Robustness, and Computational Guarantees for Sliced Wasserstein Distances", 2022 for guarantees.

How to judge? Pt.1 Max-Sliced Wasserstein

Wasserstein distances

Let $\mu, \nu \in \mathcal{P}_p$. We define

$$\mathrm{W}_p(\mu,\nu)^p := \inf_{\gamma \in C(\mu,\nu)} \int \|x-y\|^p \gamma(\mathrm{d}x, \mathrm{d}y)\,, \tag{1}$$

$$\overline{\mathrm{W}}_p(\mu,\nu) = \max_{\theta \in \mathbb{S}^{d-1}} \mathrm{W}_p(\mathfrak{p}^\theta \# \mu, \mathfrak{p}^\theta \# \nu)\,. \tag{2}$$

Empirical Wasserstein [9]

Suppose $X_{1:n} \sim \mu^{\otimes n}$, $Y_{1:n} \sim \nu^{\otimes n}$, and $\hat{\mu}_n(\mathrm{d}x) = n^{-1}\sum_{i=1}^n \delta_{X_i}(\mathrm{d}x)$ and $\hat{\nu}_n(\mathrm{d}x) = n^{-1}\sum_{i=1}^n \delta_{Y_i}(\mathrm{d}x)$. We define

$$\overline{\mathrm{W}}_{\mathrm{MC},p}(\hat{\mu}_n, \hat{\nu}_n; K) := \max_{\theta \in \Theta} \mathrm{W}_p(\mathfrak{p}^\theta \# \hat{\mu}_n, \mathfrak{p}^\theta \# \hat{\nu}_n)\,, \tag{3}$$

where $\Theta \sim \mathrm{Unif}(\mathbb{S}^{d-1})^{\otimes K}$.

[9] See Nietert et al., "Statistical, Robustness, and Computational Guarantees for Sliced Wasserstein Distances", 2022 for guarantees.

How to judge? Pt.2 Classifier two-sample test

- Generate N samples from a given sampler,
- Create a classification dataset using the test set and the generated samples.
- Split in two
- Train on one, validate on other
- Check your favorite metric.

Results: Max-Sliced Wasserstein

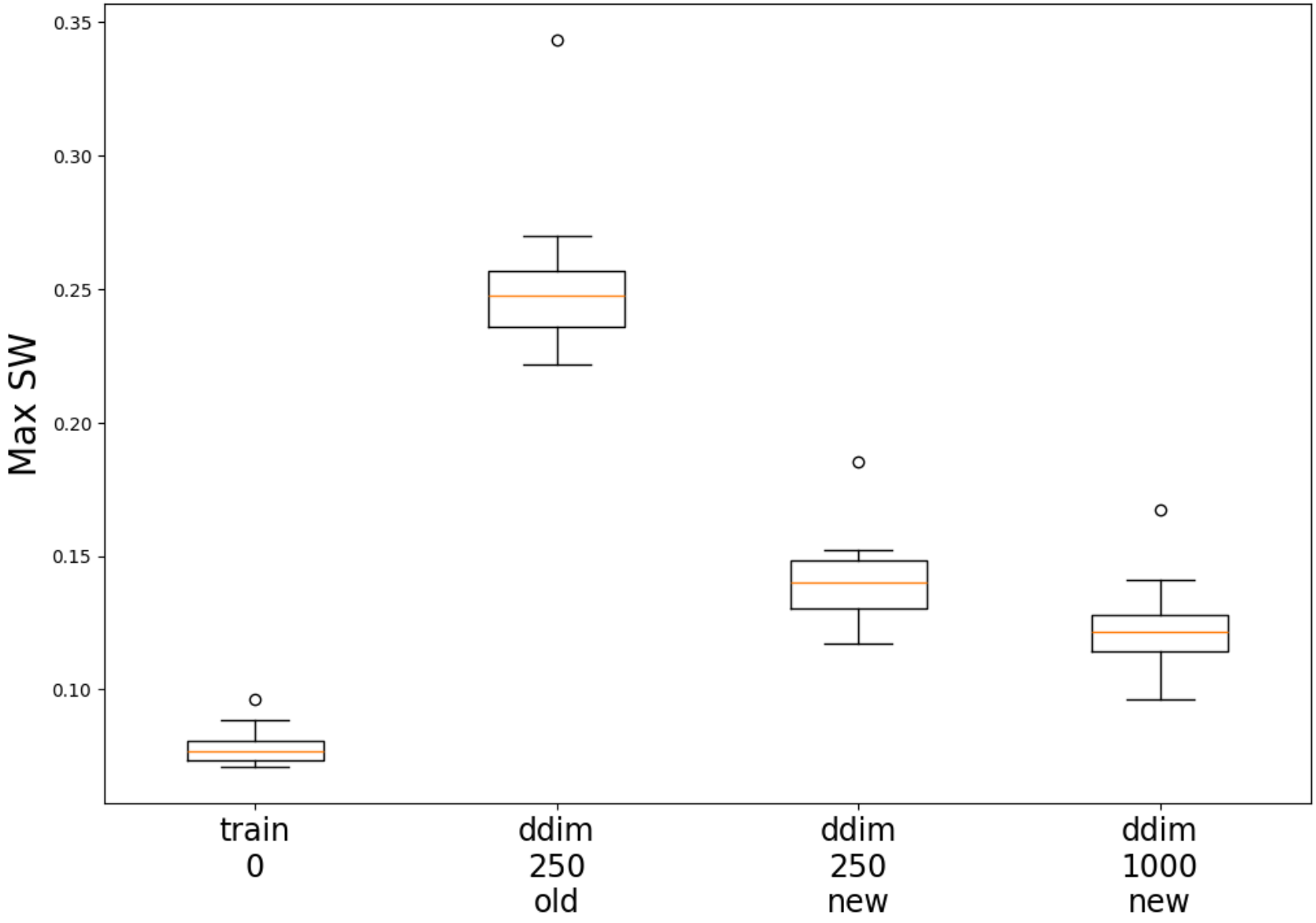


Figure: Max-Sliced Wasserstein

Results: Classifier (Resnet18)

sampler	AUC	Network
DDIM 250 old	0.84	Resnet18
DDIM 250 new	0.81	Resnet18
DDIM 1000 new	0.61	Resnet18
DDIM 1000 new	0.60	Resnet50

Table: ROC AUC.

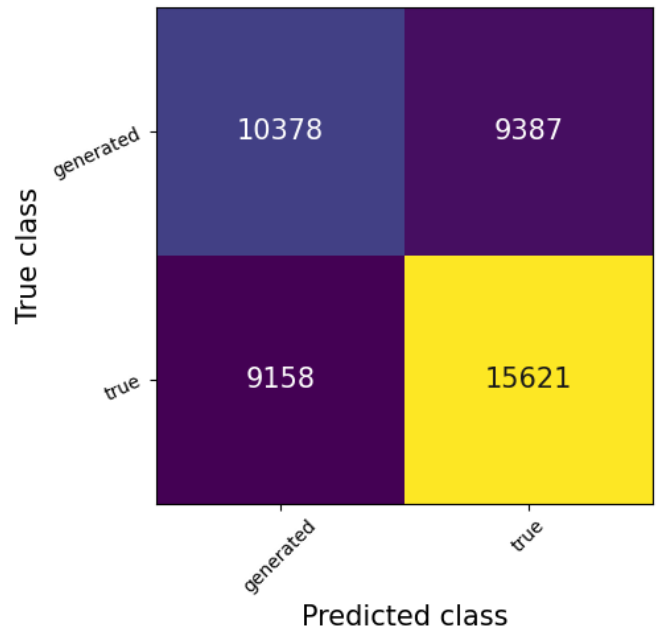
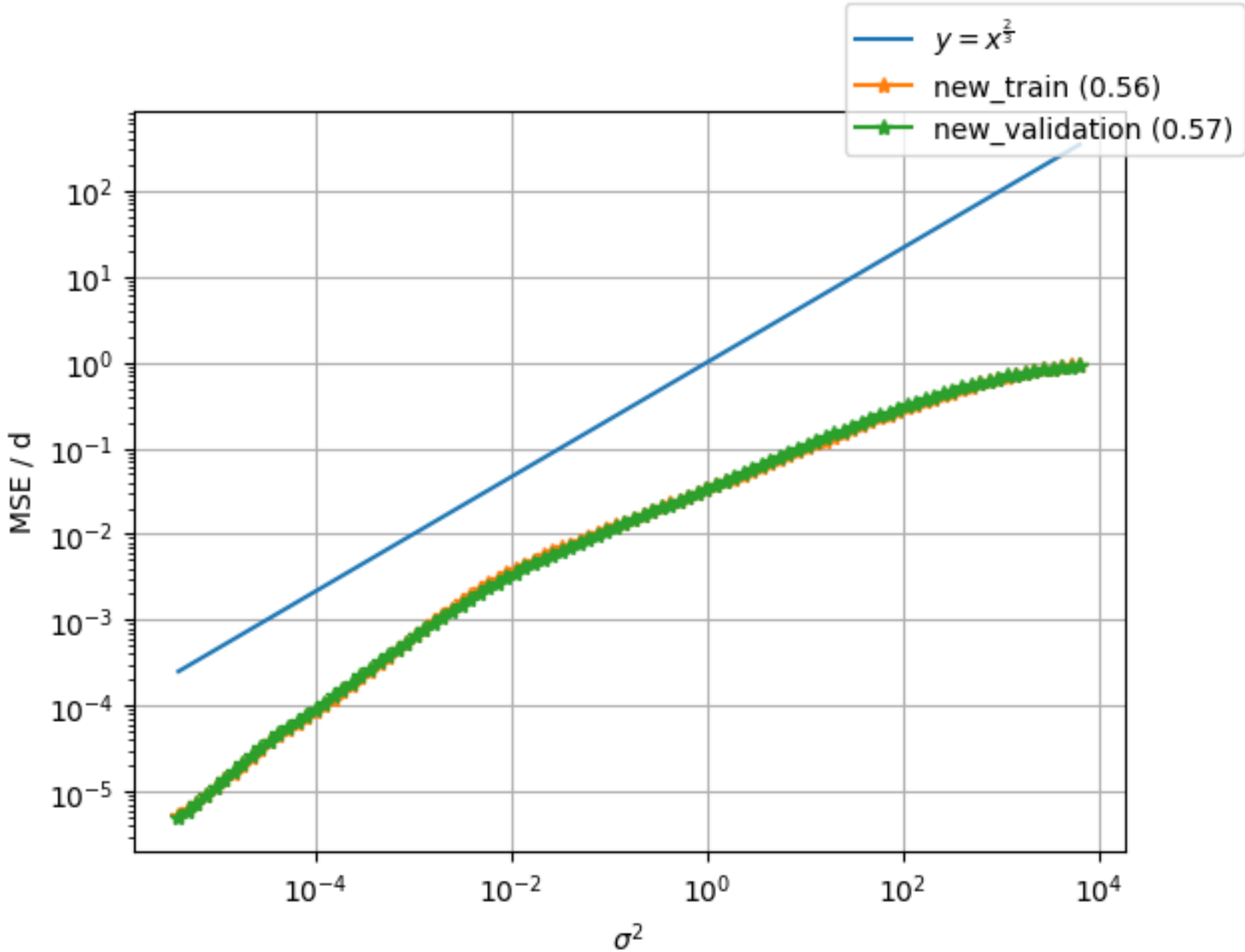


Figure: Confusion matrix for DDIM 1000 with ResNet18.

MSE vs σ^2



A glimpse of posterior sampling

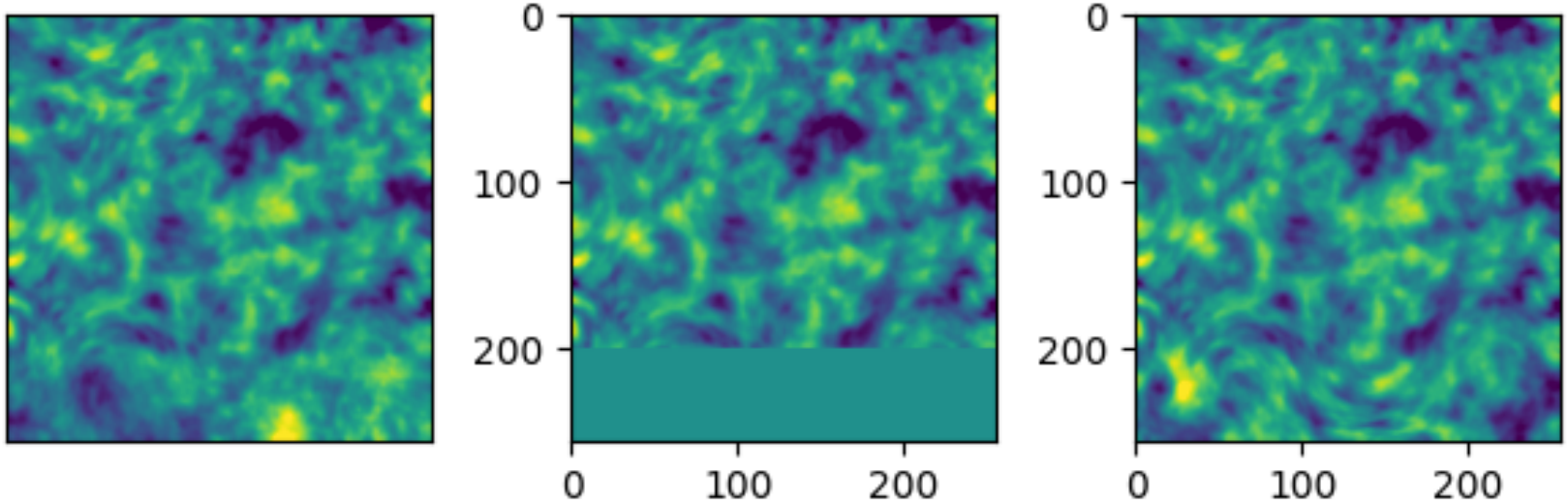


Figure: Posterior sampling result with old model.

Perspectives and challenges

- Should we treat different regularities?
- How to validate posterior sampling?