The Data Manifold under the Microscope

Controlled image manifolds for geometric analysis and manifold-fitting experiments

Marios Koulakis[1], Constantin Seibold[2]

[1] Independent Researcher
[2] Diagnostic and Interventional Radiology, University Clinic Heidelberg

ICML 2026

UNIVERSITÄTS
KLINIKUM
HEIDELBERG

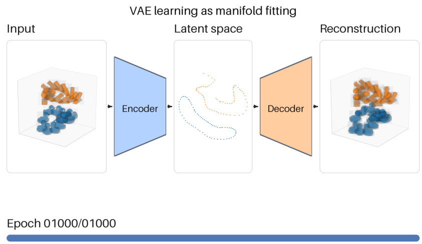
Target: 10–15 seconds.
Hi everyone, I am Marios Koulakis. In this short video, I will present our paper, *The Data Manifold under the Microscope*, where we build controlled image manifolds for studying data geometry and manifold fitting.

Motivation: Learning a data manifold

Starting point

Generative models can be viewed as procedures that fit or approximate a data manifold.

- How does approximation error scale with N?
- Which geometric quantities control this behavior?
- Can we test these questions in a controlled setting?



Target: 35 seconds.

The motivation comes from a common view of generative models. A VAE can be interpreted as indirectly fitting the manifold on which the data lives. This raises quantitative questions: how does approximation error decrease with the number of training examples, and which geometric quantities affect this behavior? To study this, we need benchmarks where the data is image-like, but the chart and geometry are not hidden from us.

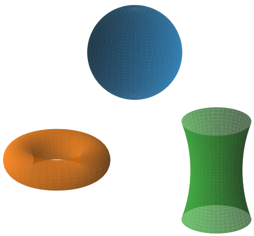
The gap: Control vs. realism

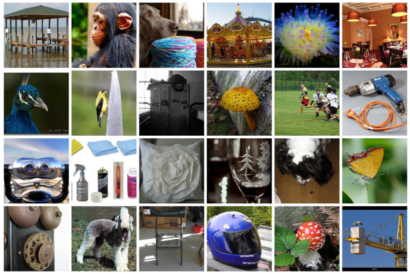
Analytic manifolds

- Known parametrization and geometry.
- Useful for validation.
- Often far from image-like data in $\mathbb{R}^{h \times w}$.

Real datasets

- Rich and realistic.
- Data-generating chart is unknown.
- Geometry is part of the learning problem.



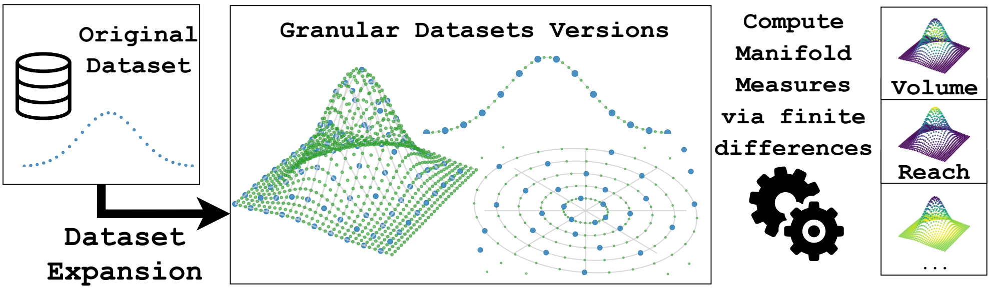


Goal: Image-like manifolds with explicit charts and measurable geometry.

Target: 35 seconds.

Existing choices leave a gap. Analytic manifolds, such as spheres, tori, or synthetic curves, are excellent for validation because their geometry is known. However, they are usually far from the ambient structure of image data. Real image datasets have the opposite problem. They are rich and relevant, but the underlying manifold is unknown. Our goal is a middle ground: image-like manifolds with explicit transformation charts and measurable geometric quantities.

Framework: controlled image manifolds



Target: 25 seconds.

For this reason, we introduced a framework that starts from existing image datasets, expands them into dense transformation grids, and then computes geometric quantities directly on the resulting image manifolds. This gives us datasets that are still image-like, but where the underlying chart and geometric measurements are accessible.

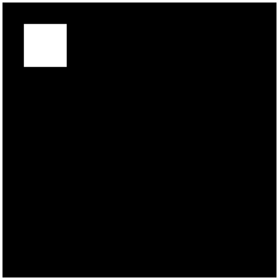
Framework: Controlled image manifolds

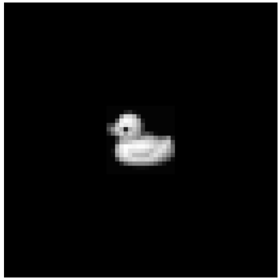
dSprites

- $d = 4$, $D = 64 \times 64 = 4096$.
- Factors: scale, orientation, x, y.
- 3 object classes.

COIL-20

- $d = 3$, $D = 4096$.
- Factors: view angle, orientation, scale.
- 20 object classes.





Dense grid in transformation space defines a discrete chart: $u : \Theta \subset \mathbb{R}^d \to \mathbb{R}^{4096}$.

Target: 45 seconds.
We do not need to start from scratch. We use two datasets from feature-disentanglement studies: dSprites and COIL-20. For dSprites, the image varies with scale, orientation, and position. For COIL-20, we use view angle and add scale and image rotation. These transformations are sampled on dense, axis-aligned grids. Therefore, the transformation variables define a discrete chart from a low-dimensional parameter space into the data manifold embedded in pixel space.

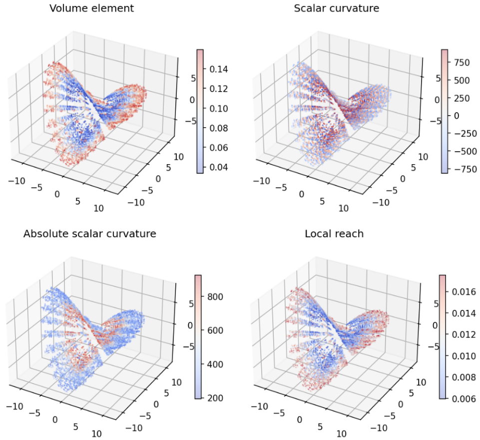
Putting geometry under the microscope

Finite differences on the grid

$$u_{,i}(x) \approx \frac{u(x + he_i) - u(x - he_i)}{2h}$$

$$g_{ij}(x) = \langle u_{,i}(x), u_{,j}(x) \rangle$$

From the metric we compute volume, scalar curvature, reach, and global summaries.



Target: 40 seconds.
Once the chart is available on a grid, we can compute geometry directly by finite differences. We estimate tangent vectors by differentiating the image-valued map with respect to each transformation coordinate. Their inner products give the Riemannian metric. From the metric, standard formulas give volume and scalar curvature. Reach is estimated from local tangent spaces. Finite differences do not solve manifold learning in general, but in this controlled setting they provide near-ground-truth measurements for testing geometric hypotheses.

Use case 1: Empirical manifold-fitting scaling

Theory gives scaling shapes

Genovese et al.:

$$C_1 n^{-\frac{2}{2+d}} \leq R_n \leq C_2 \left(\frac{\log n}{n}\right)^{\frac{2}{2+d}}$$

Fefferman et al.:

$$H(M_o, M) \lesssim C \left(\frac{\log n}{n}\right)^{\frac{1}{d}}$$

Experiment

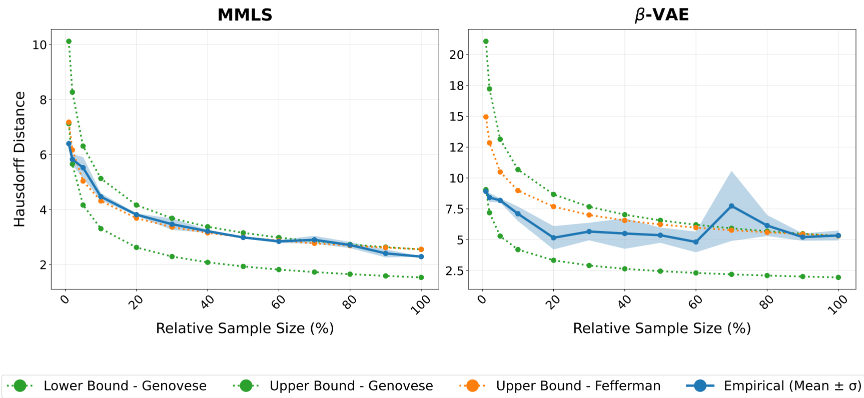
- Fit with MMLS or β-VAE.
- Vary training set size n.
- Measure directed Hausdorff proxy.
- Fit $\hat{R}_n = \hat{C} n^{-\hat{A}}$.

Target: 45 seconds.

We demonstrate the framework with two use cases. The first studies empirical scaling laws for manifold fitting. Theoretical work gives asymptotic scaling rates, including rates from Genovese et al. and Fefferman et al. In the experiments, we fit the manifold from different numbers of training samples, using either MMLS or a β-VAE. We then measure approximation error and align the theoretical scaling shapes with the empirical curves.

Scaling experiments: what the benchmark reveals

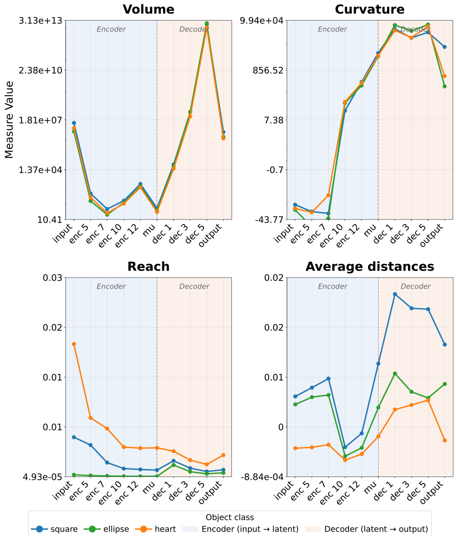
- MMLS follows a power law over most settings.
- The Fefferman-style exponent is closer to the empirical decay in these experiments.
- β-VAE curves are flatter and less well described by one power law.



Target: 45 seconds.

The benchmark makes the comparison measurable. For MMLS, the empirical curves often follow a clear power law, and in these experiments the Fefferman-style exponent is closer to the observed decay. For the β-VAE, the curves are flatter and less consistently described by one power law. This suggests that, in this configuration, the model and training objective are not acting like an efficient local manifold fitter. The broader point is that theoretical scaling claims can be confronted with finite-sample behavior when the geometry is accessible.

Use case 2: geometry across β-VAE layers



Observed trend

- **Curvature** increases.
- **Reach** decreases.
- **Class distances** increase toward the latent representation.

The model separates classes, while the intermediate manifolds become geometrically more intricate.

Target: 35 seconds.

The second use case looks inside the β-VAE. The same finite-difference machinery can be applied to hidden representations, so we can track how geometry changes across layers. In the experiments, class distances increase toward the latent representation, suggesting that the network separates the classes. At the same time, curvature increases and reach decreases, indicating that intermediate manifolds become geometrically more intricate. This attaches geometric measurements to representation learning, beyond visual inspection.

Takeaway and future directions

Takeaway

Controlled image manifolds make geometric assumptions testable.

- Explicit chart, image-like ambient space.
- Finite-difference geometry.
- Scaling and representation-analysis experiments.

Next steps: richer topologies, more datasets, and better handling of rasterization effects.

Questions and ideas are welcome at the poster.

Target: 30 seconds.
The main takeaway is that controlled image manifolds make geometric assumptions testable. The current datasets and analyses are a starting point: they provide explicit charts, image-like ambient spaces, and finite-difference geometry. Natural next steps include richer transformations such as occlusions, three-dimensional rotations, and color changes; more datasets and domains; and better treatment of topology and rasterization effects. If you have questions, suggestions or ideas, we would be happy to see you and discuss with you at our poster.