Jiongyi Wang
jiongyiwang@keio.jp

Peng Lab
Keio University
慶應義塾大学

July 18, 2025

Color Representation in CLIP

— a persistent cohomology approach —

Joint work with Peilin Zhou, Wuhan University

Introduction

- The Biological System: The brain represents an animal's position in a 2D environment using thousands of specialized neurons called grid cells. The collective firing rates of these N neurons at any given time can be seen as a point in a high-dimensional space, $\mathbb{R}^d$.
- The Emergent Structure: As the animal explores its environment, this high-dimensional point cloud of neural activity doesn't fill the space randomly. Instead, it traces out a much lower-dimensional, highly structured object—a manifold.
  - For head-direction cells (1D variable), the manifold is a ring $\mathbb{S}^1 \cong \mathbb{R}/\mathbb{Z}$.
  - For grid cells (2D variable), the manifold is a torus $\mathbb{T}^2 \cong \mathbb{R}^2/\mathbb{Z}^2$.

Neural Signature Manifold

- A neural signature manifold is a low-dimensional, highly structured object that emerges from the high-dimensional neural activity space.
- It is a topological space that captures the essential features of the neural activity, such as the firing patterns of neurons.
- The neural signature manifold can be used to understand how the brain encodes and processes information about the environment.

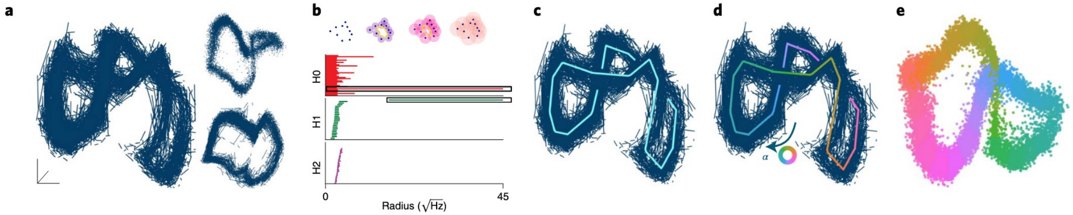


Figure: Head-direction cells in the entorhinal cortex of a rat. The firing patterns of these neurons form a ring, which is a 1D manifold in the neural activity space. Credit:[1]

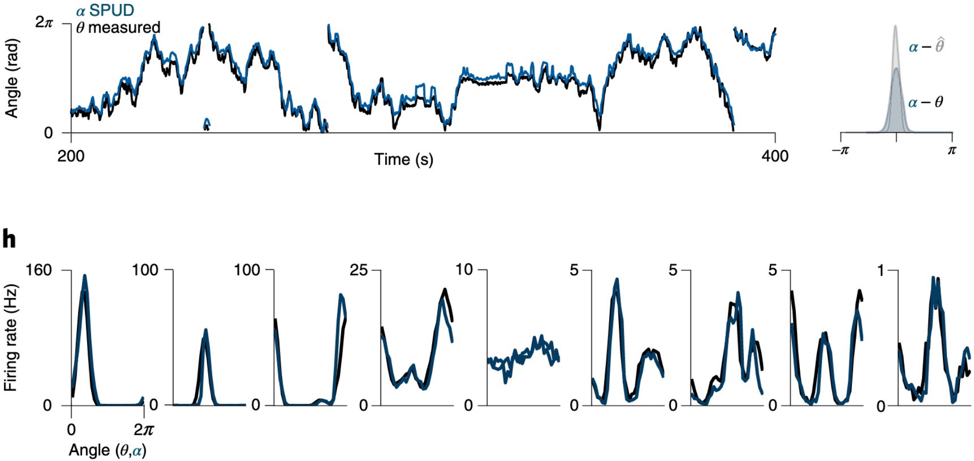


Figure: Accurately track an animal's real-time angular position from neural activity (top panel) and precisely reconstruct the corresponding tuning curves of individual neurons (bottom panel), Credit:[1]

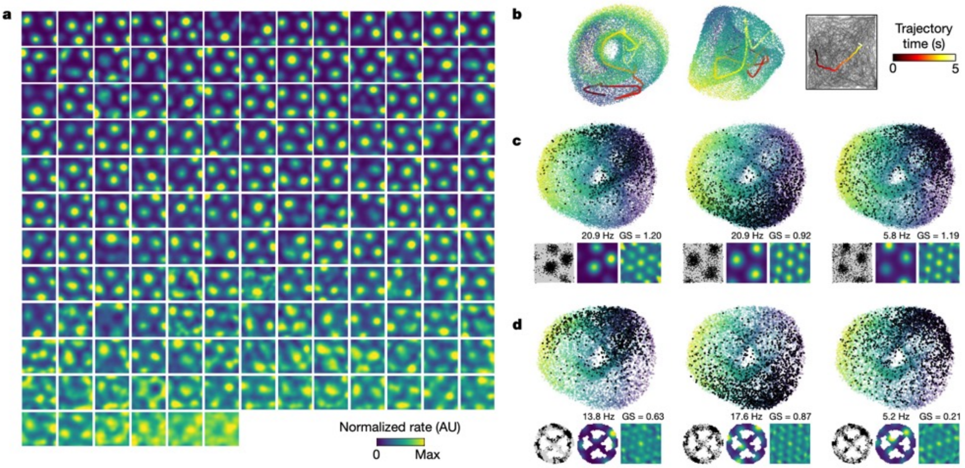


Figure: Grid cells in the entorhinal cortex of a rat. The firing patterns of these neurons form a hexagonal grid, which is a 2D manifold in the neural activity space. Credit:[2]

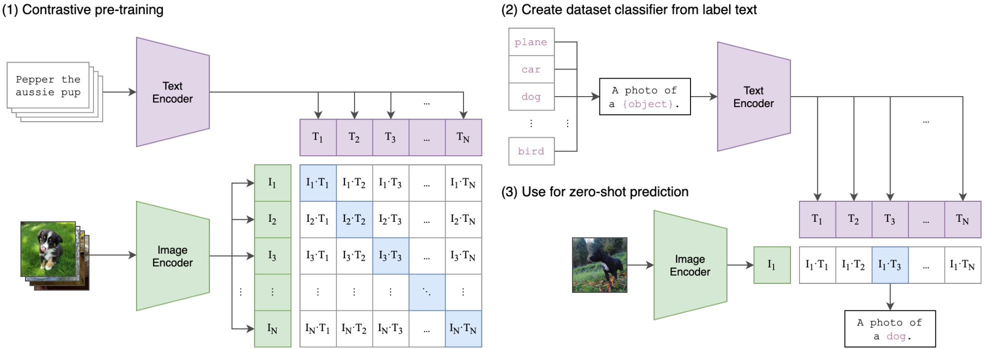
Color Organization/Comprehension in ANNs?

Questions:

1. Do Artificial Neural Networks (ANNs, aka. NNs) develop similarly organized, geometric representations for perceptual attributes like color?
2. How can we discover and verify this hidden topological structure from noisy, high-dimensional data, without prior knowledge of its shape?
3. If so, how to make use of these representations?

Contrastive Language-Image Pretraining, CLIP

- A neural network that learns visual concepts from natural language descriptions i.e., trained by (image, text) pairs.
- Can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task.
- Can be used for various tasks such as image classification, object detection, and zero-shot learning.



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$

Figure: CLIP architecture. Credit:[4]

Definition (Simplicial Complex)

Simplical complex $\mathcal{K}$ is a collection of simplices (vertices, edges, triangles, etc.) that satisfy the following properties:

- If a simplex σ is in $\mathcal{K}$, then all its faces are also in $\mathcal{K}$.
- The intersection of any two simplices in $\mathcal{K}$ is either empty or a face of both simplices.
- The empty set is also considered a simplex in $\mathcal{K}$.

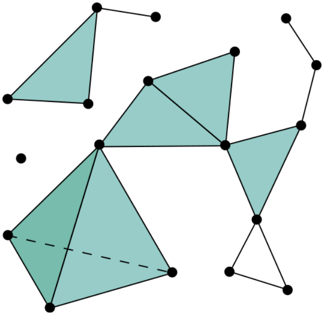


Figure: A simplicial complex $\mathcal{K}$ with vertices, edges, and triangles.

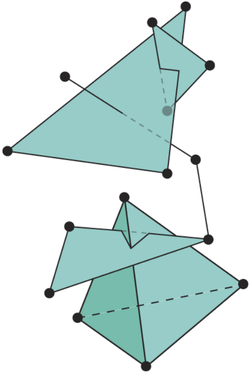


Figure: Contra example. Not a simplicial complex because the intersection of two simplices is not a face of both simplices.

Classical Example: Delaunay Triangulation

The Delaunay triangulation is a simplicial complex that captures the topological features of a set of points.

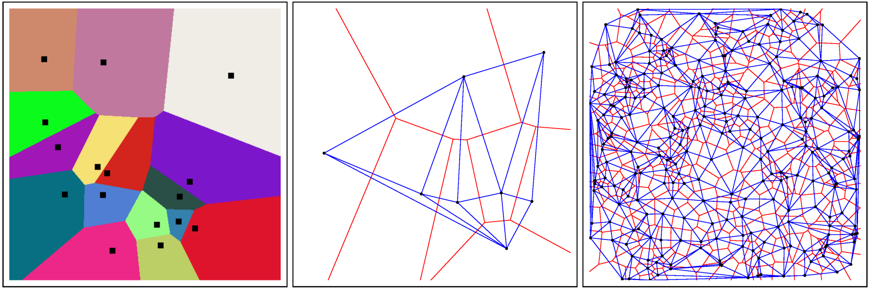


Figure: Voronoi - Delaunay triangulation in Euclidean space, which is later extended to information-geometric Cauchy manifolds by [3].

Filtration on a Dataset

Hard to discuss the topology of a dataset X directly? $\Longrightarrow$ To build an ordered-structure.

Given a point cloud (dataset) X and a scale parameter ϵ, the Vietoris-Rips (VR) Complex is defined as:

1. Vertices (0-simplices): All points in X.
2. Edges (1-simplices): An edge $\{u, v\}$ exists between points $u, v \in X$ if their distance $d(u, v) \leq \epsilon$.
3. Triangles (2-simplices): A triangle $\{u, v, w\}$ exists if all pairwise distances between u, v, w are less than or equal to ϵ.
4. ...
5. n-simplices: An $(n+1)$-set of vertices forms an n-simplex if every pair of vertices in the set is connected by an edge in the VR complex.

As ε ↑, we obtain a sequence of nested simplicial complexes, which is the filtration:

$$\mathcal{K}_0 \subseteq \mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \cdots \subseteq \mathcal{K}_m$$

where $\mathcal{K}_i = \mathcal{K}_{\varepsilon_i}$ for a sequence of scales $\varepsilon_0 < \varepsilon_1 < \varepsilon_2 < \cdots < \varepsilon_m$.

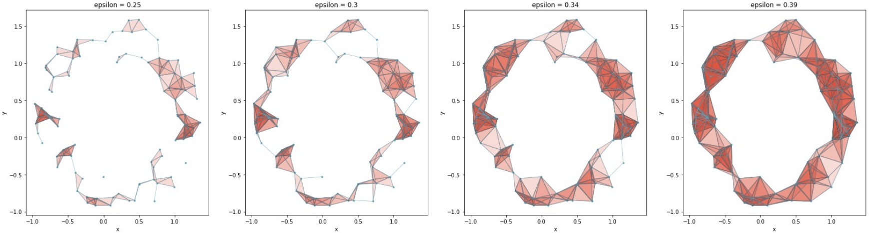


Figure: Vietoris-Rips complex filtration on a point cloud dataset. As ε increases, more simplices are added to the complex.

The birth-death of topological features

As the filtration progresses, we can track the birth and death of topological features (connected components, holes, voids, etc.):

- A feature is born when it first appears in the complex.
- A feature dies when it is no longer present in the complex.

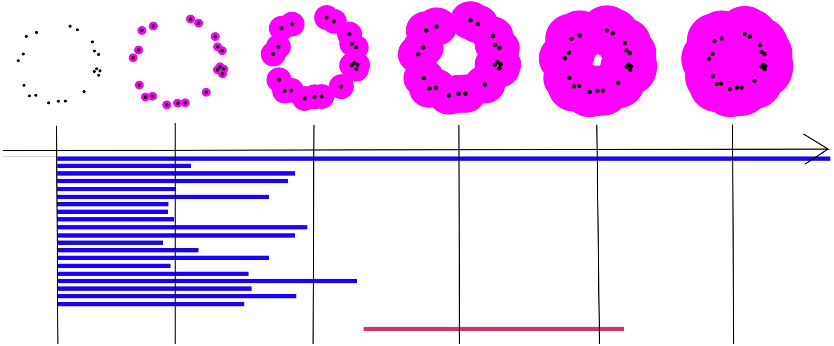


Figure: Barcode representation of the persistence diagram. Each bar represents a topological feature, with its length indicating the lifetime of the feature.

jiongyiwang@keio.jp

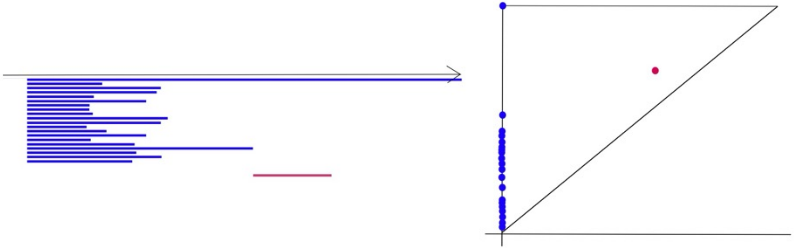


Figure: The x-axis represents the birth scale of a feature. The y-axis represents the death scale of that same feature.

Therefore, each feature is represented by a single point (birth, death). Its vertical distance from the y=x diagonal line indicates its persistence.

The birth and death of these features can be represented as a persistence diagram or a barcode.

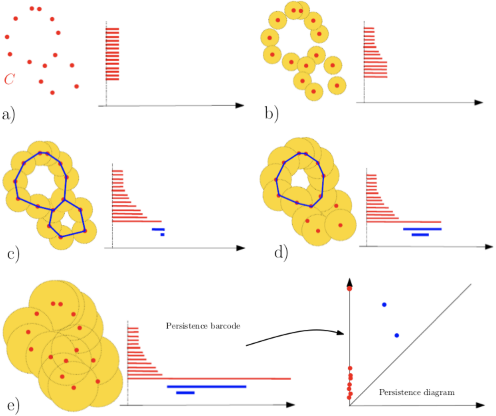


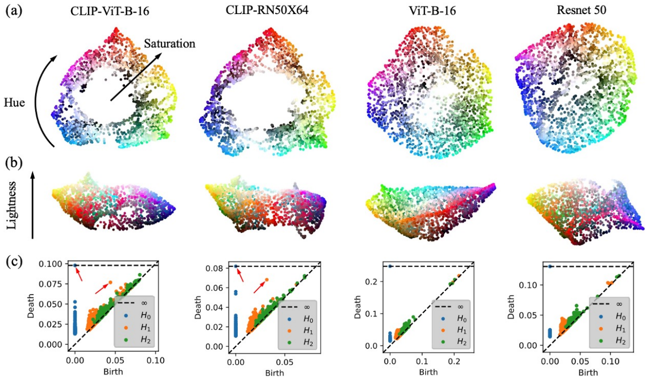
Figure: Persistence diagram showing the birth and death of topological features in a filtration.

Persistent Cohomology I

Persistent cohomology is a central tool in Topological Data Analysis (TDA). It is best understood as the algebraic dual to the more commonly known Persistent Homology. While homology tracks features like connected components and ”holes,” cohomology provides a dual perspective by tracking ”gaps” or ”cuts” across those features.

With the foundations of filtrations and duality, we can now define persistent cohomology.

The inclusion maps of the filtration, $i_j : K_j \hookrightarrow K_{j+1}$, induce linear maps between the corresponding (co)homology groups. Due to duality, their directions are opposite: Homology (Forward Map): $i_{j*} : H_p(K_j) \to H_p(K_{j+1})$ Cohomology (Backward Map): $i_j^* : H^p(K_{j+1}) \to H^p(K_j)$



The visualization of the color manifold reveals a complex structure, with colors arranged in a way that reflects their relationships in the CLIP embedding space. The edges represent the relationships between colors, and the points represent the colors themselves. This visualization provides a clear picture of how colors are organized in the CLIP embedding space, revealing the underlying structure of the color manifold.

Topology Reveals a Key Difference Between Models

Using topological data analysis, we uncovered a fundamental structural difference between how models are trained.

CLIP Models, which are trained with language supervision, form a ring-shaped color manifold with a distinct central hole.

ImageNet-trained Models, like ViT and ResNet, form a solid disk-like structure without a central hole.

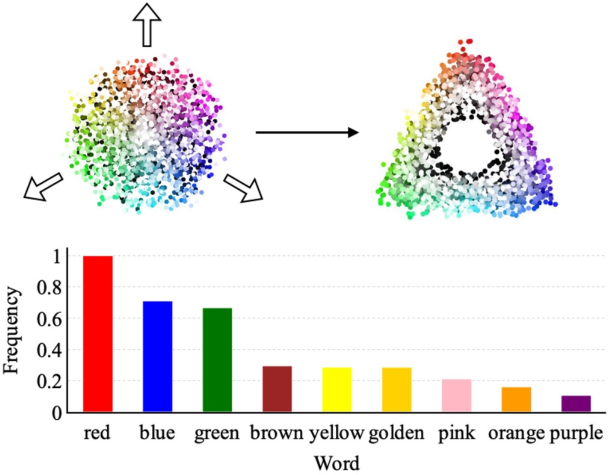
This difference is quantitatively confirmed by persistent homology, which shows a stable 1-dimensional hole (H1 feature) in CLIP models that is absent in the ImageNet models.

We hypothesize that language supervision is the primary reason for the ring structure in CLIP.

Analysis of language corpora reveals that primary color words— ”red,” ”green,” and ”blue”—appear significantly more frequently than other color terms.

This linguistic prominence appears to ”pull” or reshape the geometry of the color manifold, stretching it along these three primary color directions.

This stretching effect creates the distinctive triangular ring shape unique to CLIP, separating it from vision-only models.


Frequency
1
0.8
0.6
0.4
0.2
0
red
blue
green
brown
yellow
golden
pink
orange
purple
Word

Building on our geometric understanding, we developed two training-free methods for precise color manipulation.

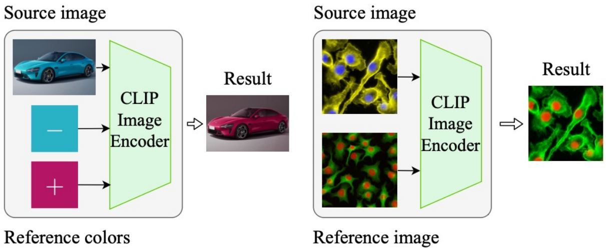
The methods directly leverage the structure of CLIP's embedding space to perform color transformations.

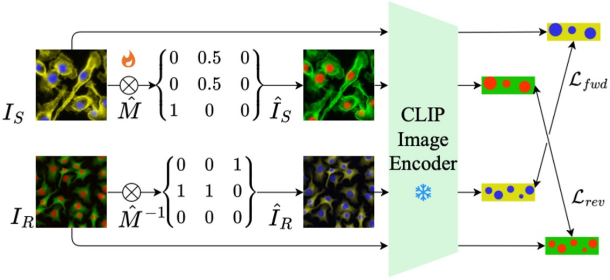
We propose:

Directional Color Transform (DCT): Manipulates specific colors by following vectors in the embedding space.

Bidirectional Color Alignment (BCA): Aligns the color style of a source image to a reference image.

Crucially, both methods only require optimizing a simple 3x3 transformation matrix and do not need any additional model training or complex architectures.


Source image
CLIP Image Encoder
Result
Reference colors
−
+
Source image
CLIP Image Encoder
Result
Reference image


I_S
$\hat{M}$
$\begin{pmatrix} 0 & 0.5 & 0 \\ 0 & 0.5 & 0 \\ 1 & 0 & 0 \end{pmatrix}$
$\hat{I}_S$
I_R
$\hat{M}^{-1}$
$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
$\hat{I}_R$
CLIP Image Encoder
$\mathcal{L}_{fwd}$
$\mathcal{L}_{rev}$

Conclusion

The research reveals that these networks arrange colors on a low-dimensional manifold similar to the HSL color space, with hues forming a circular structure

Color Transform

Thanks for your attention!

Outline

1 Heuristics

2 Uncovering the Color Manifold

3 Manipulating Colors

4 References

jiongyiwang@keio.jp

References I

[1] Rishidev Chaudhuri et al. “The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep.” In: *Nature Neuroscience* 22.9 (Sept. 2019), pp. 1512–1520.

[2] Richard Gardner et al. “Toroidal topology of population activity in grid cells.” In: *Nature* 602.7895 (Feb. 3, 2022), pp. 123–128.

[3] Frank Nielsen. “On Voronoi Diagrams on the Information-Geometric Cauchy Manifolds.” In: *Entropy* 22.7 (June 2020), p. 713.

[4] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision.” In: (Feb. 26, 2021).