Visual Anagrams Reveal Hidden Differences in Holistic Shape Processing Across Vision Models

Fenil R. Doshi, Thomas Fel, Talia Konkle, George A. Alvarez

Harvard University

Motivation

Human perception is holistic based on global shape

Thatcher Effect

Texture Change

Context Change

DNNs are emerging as de-facto models of human perception, but they remain **biased towards local information**

Critical Algorithmic Gap

Human Decision : Elephant
Model Decision : Tiger

This local bias is linked to **shortcut learning** on spurious correlations, making models **brittle and less robust**

Baker et al., 2018; Geirhos et al., 2018; Shah et al., 2020; Hermann et al., 2023

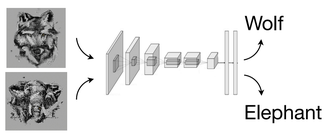
Visual Anagrams: An Absolute Test of Holistic Shape Perception

Image pairs constructed from an **identical multiset of patches** (texture-matched) rearranged to form two distinct categories

- Object Anagram Dataset = 72 pairs (expanded to 1440) from 9 objects
- Generated via multi-view diffusion from DeepFloyd

Geng et al., 2024

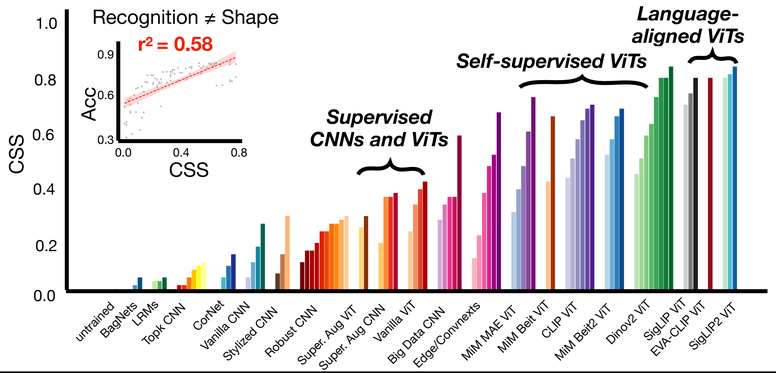
Configural Shape Score (CSS)



Model's capacity to correctly classify both anagram images measures **absolute configural sensitivity**

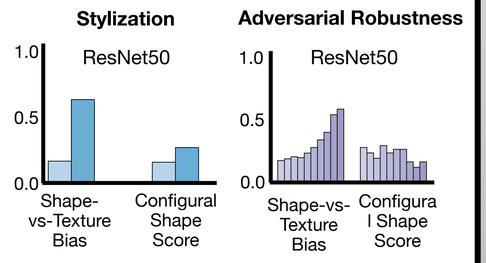
CSS = Proportion of correctly classified Anagram Pairs

Self-Supervised & Language-Aligned models dominate the CSS spectrum



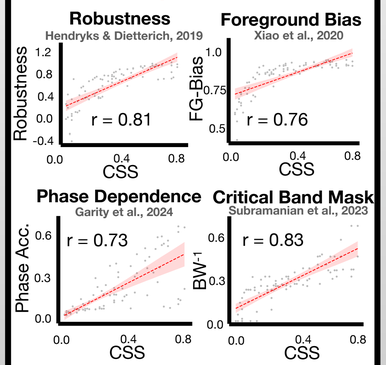
CSS dissociates from Shape-vs-Texture Bias

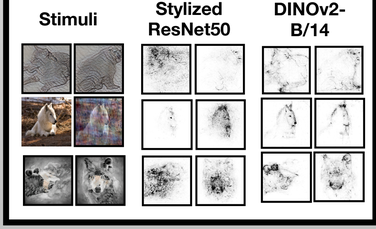
Only 41% common variance between Shape-vs-texture bias and CSS across all models



CSS → Generalization

Holistic Shape = Better Vision





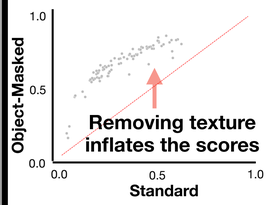
Shape-vs-Texture Bias

Standard Cue Conflict

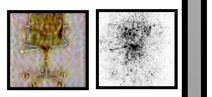
Proportion of correct decisions based on shape but this is a **relative measure** (Geirhos et al., 2018)

Object-Masked Cue Conflict

Tartaglini et al., 2022



Feature Attribution in Stylized Models

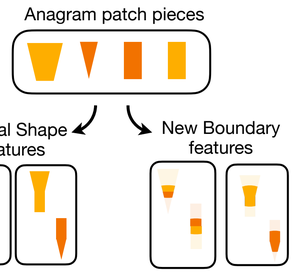


HOW TO MEASURE GLOBAL SHAPE?

We need an absolute measure of configural sensitivity, independent of local texture.

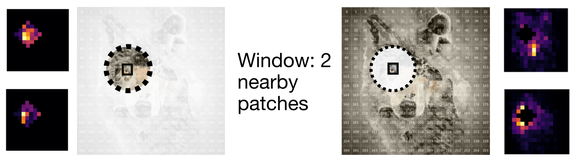
Long-Range Integration Drives Configural Processing

Could high-CSS models simply be exploiting local artifacts at the boundaries?

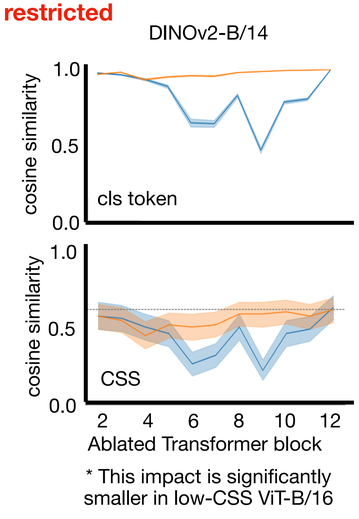


* **BagNet control rules out border-hacking,** but what about high-CSS ViTs?

To test this, we restrict attention at specific layers by forcing patches to attend locally or globally and measure impact



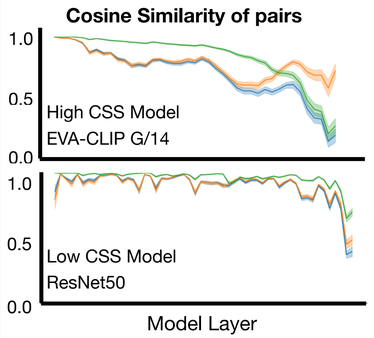
Big impact in intermediate layers when long-range attention is restricted



* This impact is significantly smaller in low-CSS ViT-B/16

Transition from Parts to Category

To generalize beyond ViTs, we track how representations evolve from local to global across all model classes



puzzle influence

$$(\text{—}) - (\text{—})$$

category influence

$$(\text{—}) - \frac{[(\text{—}) + (\text{—})]}{2}$$

r (CSS , puzzle) = -0.7

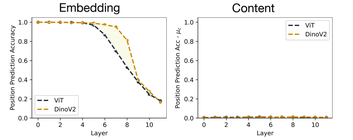
r (CSS , categ) = 0.8

High-CSS models transform local-part to global-category

Mechanism

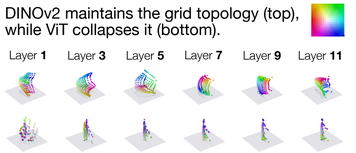
Decompose embeddings to separate position and content

Linear probes confirm that position signal is removed from content



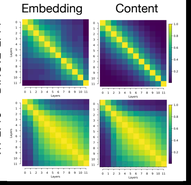
Spatial Geometry

DINOv2 maintains the grid topology (top), while ViT collapses it (bottom).



Semantics

Content undergoes Progressive Enrichment in DINOv2-B/14



* Presenting at the **MechInterp** Workshop (Dec7)