
Objective drives the consistency of representational similarity across datasets

Laure Ciernik^{1 2 3} Lorenz Linhardt^{* 1 4} Marco Morik^{* 1 4} Jonas Dippel^{1 4 5} Simon Kornblith⁶
Lukas Muttenthaler^{1 4 5}

Abstract

The Platonic Representation Hypothesis claims that recent foundation models are converging to a shared representation space as a function of their downstream task performance, irrespective of the objectives and data modalities used to train these models (Huh et al., 2024). Representational similarity is generally measured for individual datasets and is not necessarily consistent across datasets. Thus, one may wonder whether this convergence of model representations is confounded by the datasets commonly used in machine learning. Here, we propose a systematic way to measure how representational similarity between models varies with the set of stimuli used to construct the representations. We find that the objective function is a crucial factor in determining the consistency of representational similarities across datasets. Specifically, self-supervised vision models learn representations whose relative pairwise similarities generalize better from one dataset to another compared to those of image classification or image-text models. Moreover, the correspondence between representational similarities and the models' task behavior is dataset-dependent, being most strongly pronounced for single-domain datasets. Our work provides a framework for analyzing similarities of model representations across datasets and linking those similarities to differences in task behavior.

1. Introduction

Representation learning has seen remarkable progress in recent years, with state-of-the-art models achieving or even surpassing human-level performance in a wide range of computer vision tasks (Radford et al., 2021; Dehghani et al., 2023; Oquab et al., 2024; Zhai et al., 2023; Muttenthaler et al., 2024). With this progress in task performance, model representation spaces tend to approach each other irrespective of the models' training data or architecture (Huh et al., 2024). Even between models trained on different modalities, better task performances result in more similar representations. Two questions naturally arise from this phenomenon: *Do pairwise similarities of model representations transfer from one dataset to another, and are they generally linked to the models' downstream behavior?*

Investigating the correspondence between representation and behavior has a long history in representation learning and adjacent fields (e.g. Hermann & Lampinen, 2020; Sucholutsky et al., 2023). Two models can have different intermediate layer representations even though they show the same task behavior (Muttenthaler et al., 2023; Lampinen et al., 2024). However, if two models show different behavior, one can be certain that the output layer representations are different, since identical output representations map to the same behavior. What remains unclear is how behavior (e.g., downstream task accuracy) is affected by the convergence of model representations (i.e., the similarity of representations). Does this relationship depend on the similarity measure at hand, or may it be due to the nature of a dataset?

The field lacks consensus on *how to define (pairwise) representational similarity* (cf. Kriegeskorte et al., 2008; Kornblith et al., 2019a; Williams et al., 2021; Duong et al., 2023; Sucholutsky et al., 2023) and *how to systematically measure its relationship to behavior* (cf. Geirhos et al., 2021; Muttenthaler et al., 2023; Sucholutsky et al., 2023; Lampinen et al., 2024). Here, we use similarity measures that are widely used and have been empirically proven to be useful for measuring representational similarity of neural networks, such as Centered Kernel Alignment (CKA; Kornblith et al., 2019a; Raghu et al., 2021), and Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008).

^{*}Equal contribution ¹Machine Learning Group, Technische Universität Berlin, Berlin, Germany ²Hector Fellow Academy, Karlsruhe, Germany ³European Laboratory for Learning and Intelligent Systems (ELLIS), Tübingen, Germany ⁴BIFOLD—Berlin Institute for the Foundations of Learning and Data, Berlin, Germany ⁵Agnostics, Berlin, Germany ⁶Anthropic, California, United States of America. Correspondence to: Laure Ciernik <ciernik@tu-berlin.de>, Lukas Muttenthaler <lukas.muttenthaler@tu-berlin.de>.

However, pairwise representational similarity is usually measured for a single dataset and not across datasets. It may be that the representations of two vision models are highly similar for a dataset of raccoons but dissimilar for a dataset of sunflowers — due to their different *domains*. Similarly, the representations of two models may exhibit high similarity for satellite data but show low similarity for images of fruit — because of the images’ different *structure*.

Thus, it appears crucial to scrutinize the factors that determine pairwise representational similarity across sets of stimuli. Knowing about those factors will enable us to generalize pairwise representational similarity across datasets and more effectively find model sets that have learned a similar representation of the world. For a large set of diverse vision models, we examine their pairwise representational similarities for both structured and unstructured datasets from different domains. Our contributions and findings are as follows:

- We propose a principled way of measuring the consistency of pairwise similarity across datasets, by measuring whether the (relative) pairwise similarities of model representations from one dataset are transferable to the (relative) pairwise similarities of another dataset.
- We find that the objective function is a crucial factor for determining the consistency of pairwise representational similarities across datasets, whereas architecture and model size do not appear to be major factors. In particular, SSL models show more reliable generalization across stimulus sets compared to image-text and supervised models.
- Our findings show the same trend across different similarity measures, including those that emphasize local versus global representational structure.
- Pairwise similarities of model representations strongly correlate with the models’ differences in task performance for single-domain datasets, while multi-domain datasets show highly variable and specialized datasets consistently low correlations.

2. Related work

Measuring the similarity of deep neural networks. Representational similarity measures quantify the degree of overlap between the representations of two models when processing the same input. They have been used as a tool to understand deep neural networks (Li et al., 2015; Nguyen et al., 2021; Mehrer et al., 2020), investigate human-machine alignment (Sucholutsky et al., 2023; Xu & Vaziri-Pashkam, 2021), and as an objective for model distillation (Tian et al., 2020; Saha et al., 2022; Zong et al., 2023). Different representational similarity measures have been proposed, includ-

ing Canonical Correlation Analysis (CCA; Hotelling, 1992; Morcos et al., 2018), Singular Vector Canonical Correlation Analysis (SVCCA; Raghu et al., 2017), the previously mentioned CKA and RSA, among others (e.g., Williams et al., 2021; Ding et al., 2021; Barannikov et al., 2022; Cui et al., 2022). They can be distinguished from *functional similarity* measures, such as performance difference (Klabunde et al., 2025), error consistency (Geirhos et al., 2020b), and true-positive agreement (Hacohen et al., 2020), which are based on model outputs rather than internal representations (Klabunde et al., 2025), or model stitching (Bansal et al., 2021; Csiszarik et al., 2021; Merullo et al., 2023), which measures compatibility, not similarity, of internal representations (Hernandez et al., 2022).

Importantly, similarity measures may differ in the invariances they evoke and the properties of representational spaces they capture (Klabunde et al., 2025). For example, local structures may be better captured by measures based on nearest neighbors (Huh et al., 2024) or localized kernels (Kornblith et al., 2019a), while capturing global similarities may be better achieved by wide kernels. Here, we make use of kernel-based CKA, as it allows probing both local and global representation structures by varying the kernel, does not require the compared representations to be of the same dimension, and is widely used (e.g., Raghu et al., 2021; Maniparambil et al., 2024; Saha et al., 2022; Zong et al., 2023; Ding et al., 2019; Baratin et al., 2021).

Drivers of similarity in vision models. In contrast to language processing systems (e.g., Achiam et al., 2023; Team et al., 2023), vision models are trained with a wide range of different learning objectives. These include (a) supervised learning objectives using classification datasets such as ImageNet (Krizhevsky et al., 2012b); (b) SSL objectives with a reconstruction or multi-view matching loss (Chen et al., 2020a; Caron et al., 2021; He et al., 2022); and (c) objectives for jointly learning image and text representations (Radford et al., 2021; Zhai et al., 2023). This makes them interesting candidate models for analyzing representational similarities. The trend to larger model sizes (e.g. Dehghani et al., 2023) combined with increased dataset sizes (e.g. Schuhmann et al., 2022; Oquab et al., 2024) and multi-modality appears to be leading to a convergence of model representations (Huh et al., 2024). Similarity between otherwise identical networks trained from different random initializations increases with width (Morcos et al., 2018; Kornblith et al., 2019a), as predicted by theoretical studies of properties of neural networks in the infinite width limit (Lee et al., 2018; Matthews et al., 2018; Yang & Hu, 2021). Huh et al. (2024) hypothesize that convergence is further driven by task generality and inductive bias. Other works find that architecture and dataset (Raghu et al., 2021) play a central role in determining the representational structure of a model.

Consistency of similarity. Beyond the similarity function, representational comparisons depend on both the extraction layer (Kornblith et al., 2019a; Raghu et al., 2021) and the (probe) dataset (Cui et al., 2022). For example, measuring similarity for a subset of a data distribution can yield different similarities than measuring it for the whole dataset, albeit maintaining the ranking of similarities between models (Brown et al., 2024). However, recent work demonstrated that representational similarities (e.g., measured via RSA) can be used to effectively select (downstream) task-specific models (Dwivedi & Roig, 2019; Borup et al., 2023). From a behavioral perspective, it has been shown that the ranking of the *agreement* of ImageNet classifier pairs generalizes to out-of-distribution data (Baek et al., 2022). Due to the limited scope of prior work (e.g., only studying a few models (Klabunde et al., 2023; Brown et al., 2024), the driving factors of the *consistency* of representational similarities across vision datasets have not yet been identified.

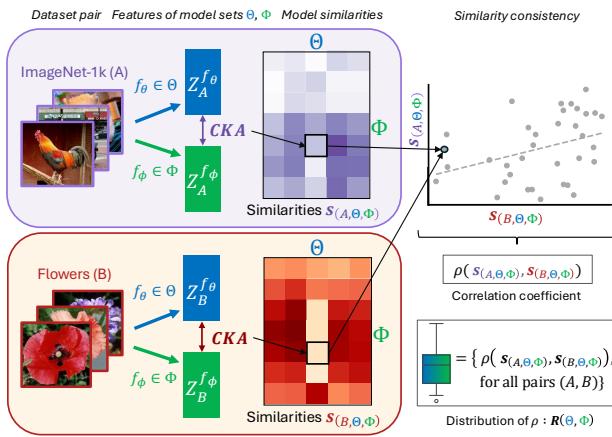


Figure 1: Pairwise similarity analysis framework. Let \mathcal{A} and \mathcal{B} be two sets of stimuli and Θ and Φ be two sets of models. For each dataset, we use all models within the two sets to extract representations, e.g., $Z_A^{f_\theta}$ being the features extracted from dataset \mathcal{A} using model f_θ . Subsequently, for each pair of models (f_θ, f_ϕ) , where $f_\theta \in \Theta$ and $f_\phi \in \Phi$, we compute the CKA similarity between their representations (i.e., $CKA(Z_A^{f_\theta}, Z_A^{f_\phi})$), yielding a similarity vector $s_{(\mathcal{A}, \Theta, \Phi)}$. This vector can be displayed as a matrix, where each entry represents the similarity between two models for a single dataset. In the scatter plot, we contrast two such vectors computed on the same model sets but evaluated on different datasets \mathcal{A} and \mathcal{B} . The Pearson correlation coefficient ρ between the similarities quantifies the consistency of similarities across the two datasets. The distribution of ρ across all dataset pairs, $R(\Theta, \Phi)$, indicates the consistency of (relative) representational similarities across stimuli.

3. Methods

In this section, we propose a novel approach for analyzing the consistency of representational similarities across multiple datasets. Fig. 1 illustrates this framework.

Extracting representations. We are interested in comparing the similarities of different vision models. We perform this comparison by extracting their latent representations and examining them. Let $f_\theta : \mathbb{R}^d \mapsto \mathbb{R}^p$ be a pretrained neural network function parameterized by a fixed set of parameters θ that maps the d -dimensional images to p -dimensional vector representations. Let $\mathcal{X} \in \mathbb{R}^{n \times d}$ be a dataset of n stacked images. For each image $x \in \mathbb{R}^d$ in a dataset, we extract its corresponding latent representations $f_\theta(x) = z \in \mathbb{R}^p$. For supervised models, we typically use the penultimate layer to extract this representation; for self-supervised vision models, the average pooling layer; and for image-text models, the image encoder. We denote by $Z \in \mathbb{R}^{n \times p}$ the matrix of stacked vector representations, which is unique for each model and dataset combination.

Computing representational similarities for a dataset. We compute the similarity between two models f_θ and f_ϕ for a dataset by applying CKA to their vector representations Z^{f_θ} and Z^{f_ϕ} . CKA computes the similarity based on the normalized Hilbert-Schmidt Independence Criterion (Gretton et al., 2005), applied to the kernel matrices of both representations. Using CKA with a linear kernel focuses on global similarity structure, while an RBF kernel with small σ measures local similarity structure (Kornblith et al., 2019a; Alvarez, 2023). We compared their behavior (shown in Fig. 3 and in Appx. F) and observed similar trends for both kernels. Therefore, we use CKA with a linear kernel for the remainder of this paper if not mentioned otherwise.

Computing the similarity requires the full kernel matrix over the dataset, which scales quadratically with the number of data points. Hence, the exact computation becomes intractable for large datasets. However, we demonstrate in Appx. C that for CKA linear a diverse subset of 10,000 samples is sufficient to accurately estimate the similarities of two models for an ImageNet-scale dataset.

Representational similarities across datasets. Because image representations depend on both a model and a dataset, the representational similarities between two models may vary across different datasets. To quantify how consistent these similarities remain across datasets (i.e., the transferability of similarity relationships), we compute the correlation coefficient for pairwise model similarities (measured for a large set of model pairs) between datasets. We can then analyze these correlations for specific subsets of models grouped by their training factors.

Specifically, we scrutinize how the pairwise similarities

of models from two model sets, Θ and Φ , correspond between two datasets, A and B . By computing the pairwise representational similarities between all model pairs (f_θ, f_ϕ), separately for each dataset, where $f_\theta \in \Theta$, $f_\phi \in \Phi$ and $f_\theta \neq f_\phi$, we obtain a similarity vector $s \in \mathbb{R}^k$, where k denotes the number of distinct model combinations. A similarity vector for two sets of models and a dataset X is then defined as $s_{(X, \Theta, \Phi)} = \left(\text{CKA}(Z_X^{f_\theta}, Z_X^{f_\phi}) \mid f_\theta \in \Theta, f_\phi \in \Phi, f_\theta \neq f_\phi \right)^\top \in \mathbb{R}^k$.

To quantify the consistency of similarities between two datasets A and B , we use the Pearson correlation coefficient between the similarity vectors $s_{(A, \Theta, \Phi)}$ and $s_{(B, \Theta, \Phi)}$, i.e., $\rho(s_{(A, \Theta, \Phi)}, s_{(B, \Theta, \Phi)})$. The Pearson correlation measures the degree to which the similarity trends between models are preserved across datasets, focusing on the relative positioning of model pairs rather than the absolute similarity values. In other words, if two models f_ϕ and f_θ are more similar than f_ψ and f_ϕ on one dataset, a high Pearson correlation coefficient indicates that this relationship transfers to the other dataset. Conversely, a low correlation value suggests that the models' relationships shift significantly depending on the dataset, showing variability in processing different data distributions. Thus, the Pearson correlation coefficient is a meaningful indicator of how dataset-dependent or dataset-invariant the model similarities are.

To assess the consistency of these correlations across dataset pairs, we examine the distribution of correlation coefficients for all possible dataset pairs within the set of available datasets \mathcal{D} : $R(\Theta, \Phi) = \{\rho(s_{(A, \Theta, \Phi)}, s_{(B, \Theta, \Phi)}) \mid A, B \in \mathcal{D}, A \neq B\}$. The spread of the correlation coefficients across all dataset pairs indicates the consistency of the relative pairwise model similarities (*similarity consistency*).

4. Experiments

Our experiments investigate the consistency of representational similarities between vision models across various datasets using our framework. Following the experimental setup (§4.1), we first analyze how model similarities transfer across datasets (§4.2, 4.4) and find consistent model groups through similarity clustering (§4.3). We then investigate how model and dataset characteristics influence similarity consistency (§4.5 and 4.6), and conclude by examining the relationship between representational similarities and models' performance differences on downstream tasks (§4.7). The code and the data to run our analyses and reproduce the experimental results are publicly available at https://github.com/lciernik/similarity_consistency.

4.1. Setup

Models. We evaluated 64 vision models, representing a diverse range of architectures, training paradigms, and parameter scales. We focused on general-purpose vision models trained on large-scale datasets (minimum ImageNet-1k size) with broad semantic diversity, excluding models trained on specialized or synthetic datasets. A complete list of models and their characteristics can be found in Appx. A.

Model grouping. Instead of focusing on the representational similarities of individual model pairs, we analyze aggregated similarities between sets of models. We categorize models based on four attributes: *training objective*, *training data (size)*, *model architecture*, and *model size* (see Tab. 3). This allows us to examine how different model characteristics relate to the behavior of representational similarities while acknowledging that these attributes are often correlated across models (see Fig. 9).

Datasets. We evaluated pairwise model similarities across 20 datasets from the CLIP benchmark (Cherti & Beaumont, 2022) and 3 datasets from Breeds (Santurkar et al., 2021). This set includes various VTAB datasets as well as ImageNet-1k. Following the categorization proposed in (Zhai et al., 2020), we classified the datasets into three main types: natural (e.g., ImageNet-1k), specialized (e.g., PCAM), and structured (e.g., DTD) image datasets. Furthermore, we partition the natural image datasets into single- and multi-domain categories. A list of all datasets can be found in Tab. 1.

4.2. Do representational similarities transfer across datasets?

The transferability of model similarities across diverse datasets is crucial for model comparison and selection. If similarities are consistent across datasets, we can evaluate models on a single dataset and generalize their relationships to other datasets. We analyzed this by computing pairwise CKA similarities for all model pairs across our datasets.

Model similarities (e.g., the range of values and the similarity trends) appear to vary across datasets (Fig. 2, Appx. D). We observe the highest similarity within image-text models (yellow boxes), while a group of self-supervised models (white boxes) shows consistent similarity trends across datasets. Yet, there are also model groups with higher variability (cyan boxes), mainly containing supervised models. The standard deviation matrix (rightmost matrix in Fig. 2) quantifies these differences, indicating that model similarities do not transfer uniformly across datasets.

To further investigate the transferability of model similarities across datasets, we examine the relationship between the mean and standard deviation of similarity values. The left panel in Fig. 3 shows an inverse U-shape trend (fur-

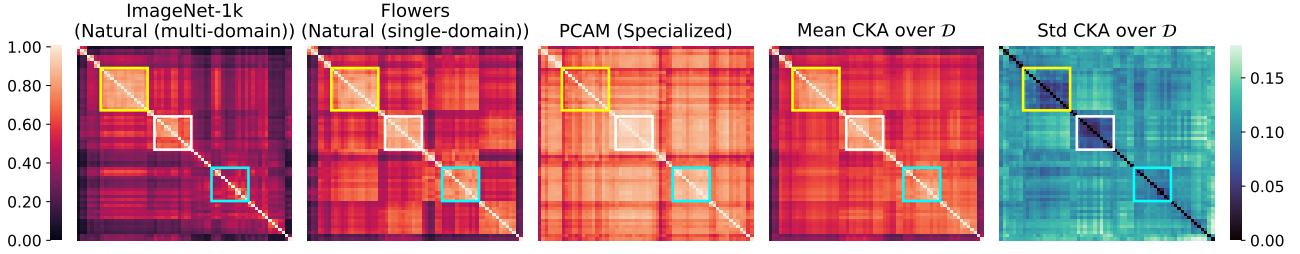


Figure 2: Representational similarity using linear CKA. Left to right: natural multi- and single-domain, and specialized datasets, followed by mean and standard deviation across all datasets. Models (rows and columns) are ordered by a *hierarchical clustering* of the mean matrix. Yellow and white boxes highlight regions with more stable similarity patterns across datasets, corresponding to some image-text (yellow) and self-supervised model pairs (white), while cyan boxes show higher variability for mainly supervised model pairs.

ther analyzed in Appx. E): model pairs with extreme mean similarities exhibit low variability, while variability for mid-range similarities is higher. To verify that this observed variability is not an artifact of the similarity metric, we compare the mean similarity values obtained using linear CKA with two other metrics: CKA with an RBF kernel where $\sigma = 0.2$ (a local similarity measure, cf., Appx. F for σ selection) and RSA using Spearman correlation (a global similarity measure). The high correlations among these measures confirm that mid-range similarity values consistently vary across metrics, indicating limited stability across datasets.

Given this variability, we investigate persistent trends in model relationships across datasets by addressing two questions: a) Can we identify subgroups of models whose similarity to other models (consistently) cluster across different datasets? b) Do the *relative* representational similarities within or between these subgroups remain stable/consistent across datasets? Answering these questions will allow us to make generalizable, albeit potentially weaker, claims about models’ relationships from a single data source.

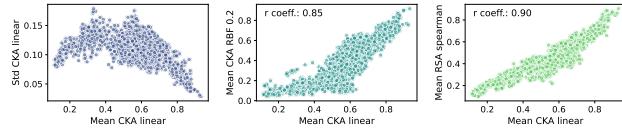


Figure 3: Variability of representational similarities across datasets. **Left:** Mean versus standard deviation of pairwise similarities. **Center/Right:** Mean similarity values using CKA RBF ($\sigma = 0.2$) or RSA as a function of linear CKA. High Pearson correlation coefficients indicate consistency across different similarity metrics.

4.3. Do representational similarities cluster according to model categories?

We qualitatively assess the alignment between our predefined model categories (§4.1) and the clustering patterns

observed in the t-SNE embedding of the model-model similarity matrices, as illustrated in Fig. 4. The training objective appears to be the most distinctive categorization, with all image-text and most SSL models forming distinct clusters across all three datasets. The objective and dataset categories show substantial overlap, potentially confounding further analysis. For instance, only image-text models have been trained on XLarge datasets, while all models trained on ImageNet-21k are supervised models (Appx. A). Additionally, no image-text model has been trained on ImageNet-1k or ImageNet-21k. In contrast, no obvious clustering is discernible based on architecture or model size. These observations provide initial evidence that the training objective might play a key role in representational similarity.

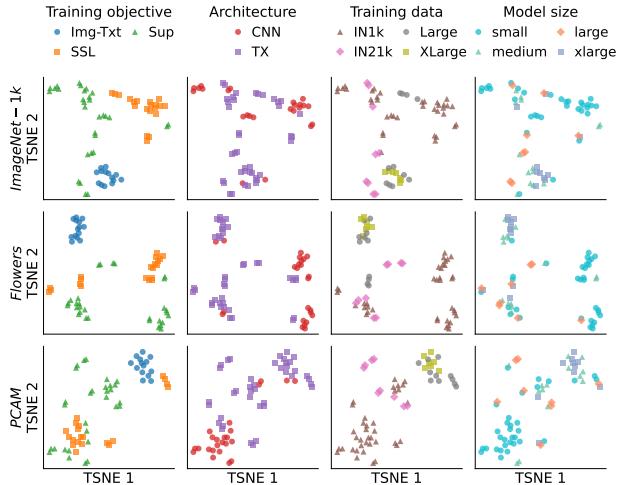


Figure 4: t-SNE visualization of model similarity matrices for ImageNet-1k, Flowers, and PatchCamelyon (PCAM) datasets. Embeddings are color-coded by model attributes: training objective, architecture, training data, and model size.

4.4. Do relative representational similarities remain consistent across different datasets?

Given the observation that representational similarities are not directly transferable across datasets but depend on both the models and the dataset, we now examine the consistency of *relative* representational similarities across dataset pairs. Following the method described in §3, the top row of Fig. 5 reveals positive correlations between pairwise model similarities across three exemplary dataset pairs. Across all dataset pairs, the mean Pearson correlation coefficient is 0.756 (std=0.124). This indicates that the ordering of pairwise model similarities is largely consistent across datasets, suggesting a degree of transferability in the relationships of representational similarities. However, the high standard deviation and the distributions of similarities shown in Fig. 5 suggest the existence of model-pair groups that may exhibit distinct consistencies of similarity values.

The second row in Fig. 5 shows model pairs within the same training objective, highlighting differences in group-specific consistency levels. We find that SSL models show the strongest similarity consistency across datasets, image-text models form a cluster of generally high similarities, and supervised models achieve the weakest consistency. In the following section, we will analyze the aggregated measure across all dataset pairs and revisit all categories in detail.

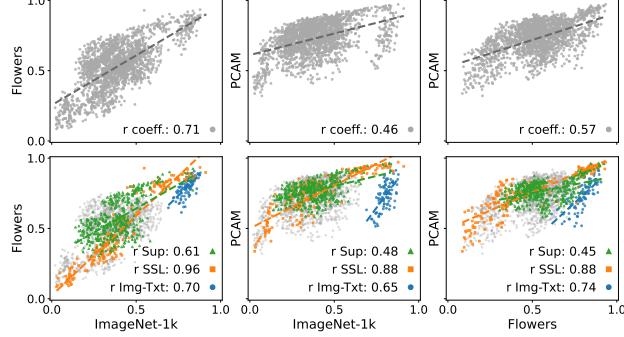


Figure 5: Pairwise representational similarities for model pairs across three dataset combinations: ImageNet-1k vs. Flowers, ImageNet-1k vs. PCAM, and Flowers vs. PCAM.

Top row: Pearson correlation coefficients for all model pairs. **Bottom row:** Same data, with colored points highlighting model pairs within the same training objective category ($\Phi_{\text{Img-Txt}}$, Φ_{SSL} , and Φ_{Sup}).

4.5. Which model categories influence similarity consistency?

Fig. 6 shows the distribution of Pearson correlation coefficients $R(\Theta, \Phi)$ for each combination of model sets (Θ, Φ) across all dataset pairs.

Training objective. Self-supervised models (Φ_{SSL}) show

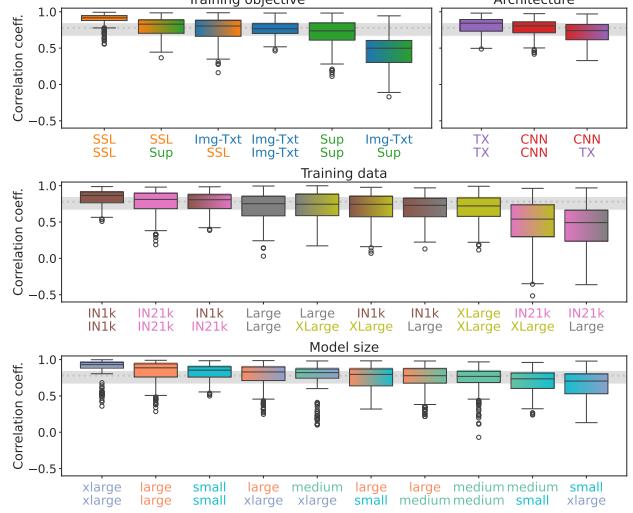


Figure 6: Distribution of correlation coefficients for model similarities across dataset pairs (consistency values). Each subplot represents a model category: training objective, architecture, training data, and model size. Within each category, distributions show the similarity consistencies between model set combinations (e.g., $R(\Phi_{\text{Img-Txt}}, \Phi_{\text{Sup}})$). The boxes in each subplot are sorted in decreasing median correlation. The dotted line indicates the overall median, while the gray area spans the 25th to 75th percentiles of correlations across all model pairs (no category consideration).

the highest similarity consistency, even when compared to models with different training objectives. This suggests that their relative similarities are most transferable across datasets, likely due to the ability of SSL to capture dataset-independent features. In contrast, supervised models show the weakest correlations, particularly when compared to image-text models ($\Phi_{\text{Img-Txt}}$, Φ_{Sup}), indicating a less reliable transfer of relative pairwise similarities across datasets.

Model architecture. While no significant differences in similarity consistency exist between transformers and convolutional networks, comparisons across architectures show slightly lower consistency, potentially indicating distinct architectural inductive biases.

Training data. We find a low median correlation and a high variance for the model set pairs ($\Phi_{\text{IN21k}}, \Phi_{\text{Large}}$) and ($\Phi_{\text{IN21k}}, \Phi_{\text{XLarge}}$). However, these effects may be confounded by the training objectives, as both Φ_{Large} and Φ_{XLarge} are mainly image-text models, while Φ_{IN21k} consists of supervised models. On the other hand, Φ_{IN1k} , which includes models trained on an ImageNet-21k-like dataset but with diverse training objectives, does not show the same trend. Consequently, training data (size) appears to have a less significant impact on the consistency of similarities than other factors. However, this applies to models trained on large and diverse datasets; as shown in Appx. I, training

on less diverse data noticeably affects consistency.

Model size. Models in the same size category show a higher similarity consistency than models of different size categories (e.g., Φ_{small} and Φ_{large}). This indicates model size influences the transferability of relative similarities across datasets, though less significantly than training objective.

Our analysis identifies the training objective as a primary factor influencing similarity consistency, with model size playing a less pronounced role. Self-supervised models consistently demonstrate above-average correlations across datasets. The effects of model architecture and training data (when mainly considering dataset size) are minimal, suggesting that these factors are less critical in determining the transferability of relative similarities across datasets (see Appx. G for an analysis using single-element model sets controlling for model architecture and size).

4.6. Do dataset categories influence similarity consistency?

Previously, we analyzed similarity consistency from the perspective of models, noting substantial variation in consistency across all dataset pairs. In this section, we shift our focus to identifying groups of datasets that exhibit more stable relative similarities, aiming to uncover variables that might explain these variations.

The top panel in Fig. 7 shows the Pearson correlations of all model pairs for each dataset pair individually, i.e., $\rho(s_{(A, \Phi_{\text{All}}, \Phi_{\text{All}})}, s_{(B, \Phi_{\text{All}}, \Phi_{\text{All}})})$ for all datasets pairs (A, B) . The consistency of representational similarities varies strongly across dataset pairs, showing higher consistency within dataset categories than across them, as exemplified by high consistencies within natural multi-domain datasets but lower ones when paired with specialized datasets. We hypothesize that high-consistency dataset categories largely share visual features. For example, the Breeds datasets, Caltech-101, Country-211, STL-10, and Pets contain (often-times centered) natural objects and scenes. For CIFAR-10 and CIFAR-100, the particularly high consistency may also be supported by their relatively low resolution. Moreover, some datasets are inconsistent with almost any other dataset, particularly FGVC Aircraft and medical imaging datasets (Diabetic Retinopathy and PCAM) – most likely due to their unique visual structure, which differs across medical domains. Interestingly, ImageNet-1k exhibits a milder yet significant pattern of inconsistency, in particular, in combination with CIFAR-10 and CIFAR-100.

To investigate whether specific dataset categories are responsible for the difference in consistency observed in the last section, we focus on the correlations for the model sets with the highest (Φ_{SSL} , Φ_{SSL}) and lowest ($\Phi_{\text{Img-Txt}}$, Φ_{Sup}) consistency (see the bottom left and right panels of Fig. 7). The

two matrices reveal distinctly different patterns, highlighting how model categories influence similarity consistency.

For SSL model pairs, we generally observe high consistency across most datasets, with FGVC Aircraft being an exception, showing lower correlations with the rest. This confirms that SSL models tend to produce more transferable similarity structures across datasets. In contrast, image-text/supervised model pairs show relatively higher consistency for multi-domain than single-domain datasets, suggesting better similarity transfer when semantic categories in the dataset match the training data of the models. However, Pascal VOC 2007 appears as an outlier when compared to ImageNet-1k and its subsets, likely because its bounding-box-cropped images lack contextual information, causing single-domain-like behavior.

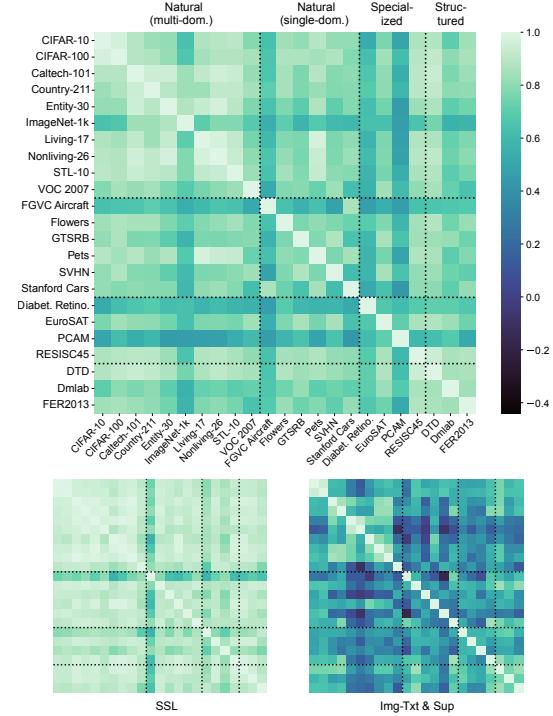


Figure 7: Pearson correlation of representational similarities for all dataset pairs computed on different model sets. **Top:** all available model pairs; **Bottom left:** SSL models (Φ_{SSL} , Φ_{SSL}); **Bottom right:** image-text and supervised models ($\Phi_{\text{Img-Txt}}$, Φ_{Sup}). Dataset categories are delineated by dotted lines.

Overall, we observe the highest consistency within multi-domain image datasets, suggesting that models are consistent when dealing with diverse semantic categories. In contrast, we find weaker consistency for ($\Phi_{\text{Img-Txt}}$, Φ_{Sup}) between multi-domain and single-domain datasets, indicating that even though these models effectively generalize across diverse datasets, their representational structures differ substantially when evaluated on single domains. Self-

supervised models do not exhibit this sensitivity to datasets; their similarity relationships appear to be more stimulus-invariant, maintaining consistent similarities regardless of whether the data is multi-domain or single-domain.

4.7. Can representational similarity predict performance gaps?

To investigate the relationship between downstream task performance and representational similarity, we contrast two measures for each dataset: (1) the CKA value for each model pair and (2) the absolute difference in classification performance, referred to as *performance gap* (see Appx. B for details on performance computation). Fig. 8 shows this relationship for one dataset per category (see Fig. 21 & 22 for additional datasets). While prior work suggests well-performing models produce similar representations (Huh et al., 2024), we find this relationship is dataset-dependent.

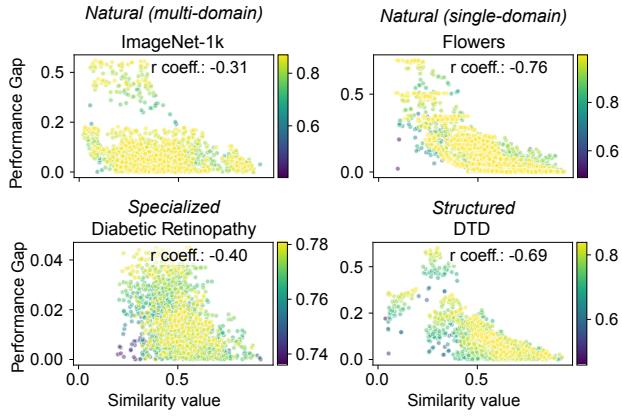


Figure 8: Model similarity (CKA linear) versus absolute difference in downstream task performance (top-1 accuracy) for model pairs across four dataset categories: natural multi-domain, natural single-domain, specialized, and structured. The color of each point indicates the downstream task accuracy of the better-performing model.

Natural multi-domain datasets, such as ImageNet-1k, show weak negative correlation between similarity and performance gap. Many model pairs achieve high performance despite low representational similarity (yellow dots in the lower left corner). This suggests that multi-domain datasets might provide richer contextual information, enabling multiple high-performance strategies that do not converge to similar representations. However, correlation coefficients in this dataset category show notable variability (see Appx. J).

Natural single-domain and Structured datasets, in contrast, exhibit a strong negative correlation between similarity and performance gap. The absence of high-performing, dissimilar model pairs suggests a close link between performance and representational similarity, indicating that

successful models in these datasets rely on capturing specific discriminative features for closely related classes. Observing this behavior in both structured and single-domain datasets suggests that models may depend on representing more structural attributes to distinguish between classes. However, the limited number of structured datasets restricts broader conclusions.

In summary, our analysis shows different trends depending on the dataset. Multi-domain datasets support a wider range of successful strategies, while performing well on single-domain and structured datasets appears to require a more limited set of features, resulting in stronger correlations between performance and similarity. This phenomenon aligns with findings from previous studies, which have shown that models can leverage contextual cues in multi-domain datasets to achieve strong performance without developing robust, generalizing features (Lapuschkin et al., 2019; Geirhos et al., 2019; 2020a).

5. Discussion

The representations of foundation models appear to converge to a canonical representation, irrespective of their training data and objectives (Huh et al., 2024). However, this may just be an artifact of the datasets that the community commonly uses for evaluation. Thus, in this paper, we presented a *systematic way of analyzing pairwise similarities of model representations across sets of stimuli*. This approach allowed us to examine whether representational similarities generalize across datasets and identify variables influencing their consistency.

We found that the *training objective* of models is a central factor for determining the consistency of pairwise representational similarities across datasets, whereas architecture and model size appear less important. Among training paradigms, image-text training leads to the most similar representation spaces, irrespective of the model’s training data or architecture. For individual datasets, SSL model representations are not as similar among themselves as the representations of image-text models, but the pairwise similarities between SSL models and other models are highly consistent across different datasets.

We hypothesize that both image-text and supervised models learn distinct semantic categories during training that lead to representations whose features overfit to these categories and, thus, do not generalize well to datasets that do not contain them. In contrast, self-supervised pure vision models, which are not constrained to learn explicit semantic categories during training, may develop representations that better accommodate diverse stimuli. This leads to relative pairwise similarities that vary less across datasets than the similarities of image-text or supervised model representa-

tions.

Alternatively, the SSL model representations, usually extracted from the average pooling layer (cf., Chen et al., 2020b; Muttenthaler et al., 2023; Tian et al., 2024), may be further away from the models’ task behavior than the image encoder and penultimate layer representations from image-text and supervised models. One benefit of our framework is that it can be applied to the representations of any model layer, allowing future work to test this hypothesis. Furthermore, in contrast to image-text and supervised models, the tasks used to train SSL models are extremely heterogeneous. This may yet be another reason for the higher consistency of representational similarities among SSL models.

Finally, we have shown that the pairwise similarities between model representations are only in part predictive of their differences in downstream task behavior. While we observe a strong correlation on datasets with limited contextual information, for others, e.g., ImageNet-1k, there is no obvious correspondence between the pairwise similarities of model representations and the models’ differences in task performance. Moreover, the relationship between pairwise similarities and task performance is substantially different between natural multi-domain datasets. This is surprising in light of recent findings that have shown a convergence of model representations as a function of task performance (Huh et al., 2024) and a strong linear relationship between a model’s ImageNet performance (i.e., performance on the training set distribution) and its downstream accuracy (Kornblith et al., 2019b). This suggests that the transferability of pairwise similarities across datasets follows a more complex relationship than the transferability of task performance. While different objectives and hyperparameters lead to similar task performances (Kornblith et al., 2021), they seem to change the similarity spaces of the learned representations in a way that affects the relationship between pairwise model similarities and differences in downstream behavior between datasets. Therefore, whether the claims of The Platonic Representation Hypothesis hold or not depends on the nature of a dataset rather than being true for all sets of stimuli.

Conclusion. In summary, we presented a blueprint for systematically analyzing pairwise representational similarities across different sets of stimuli. This allowed us to demonstrate that *pairwise similarities rarely transfer from one dataset to another and depend on both the models’ objective functions and the datasets’ structure and domain*. The community has long optimized for finding the optimal hyperparameters to improve task performance (e.g. Tolstikhin et al., 2021; Steiner et al., 2022; Liu et al., 2022; Dehghani et al., 2023) but neglected its ramifications for the similarity structure of the models’ representation spaces. Only few recent works investigated how representations are affected

by training tasks (e.g. Lampinen et al., 2024; Kornblith et al., 2021). We believe that pinpointing the variables of model training affecting (dis-)similarities between model representations beyond the training dataset will ultimately help to build more interpretable systems with which humans are more likely to interact (Lake & Baroni, 2023; Sucholutsky et al., 2023; Muttenthaler et al., 2024; Tessler et al., 2024).

Acknowledgments

LC is funded by the Hector Fellow Academy and is grateful for their support.

Impact Statement

This work investigates the consistency of representational similarities across datasets and provides new insights into how stimuli and model characteristics interact. Our findings challenge assumptions about universal representational convergence and highlight the nuanced relationship between representations and downstream task performance, paving the way for more robust and aligned models. While these findings contribute to the advancement of machine learning, we do not foresee any immediate negative societal consequences requiring specific emphasis.

References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Alvarez, S. A. Gaussian rbf centered kernel alignment (cka) in the large-bandwidth limit. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6587–6593, 2023. doi: 10.1109/TPAMI.2022.3216518.
- Baek, C., Jiang, Y., Raghunathan, A., and Kolter, J. Z. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 19274–19289. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7a8d388b7a17df480856dff1cc079b08-Paper-Conference.pdf.
- Bansal, Y., Nakkiran, P., and Barak, B. Revisiting model stitching to compare neural representations. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7a8d388b7a17df480856dff1cc079b08-Paper-Conference.pdf.

- s.cc/paper_files/paper/2021/file/01d-ed4259d101feb739b06c399e9cd9c-Paper.pdf.
- Bao, H., Dong, L., Piao, S., and Wei, F. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=p-BhZSz59o4>.
- Barannikov, S., Trofimov, I., Balabin, N., and Burnaev, E. Representation topology divergence: A method for comparing neural network representations. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1607–1626. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/barannikov22a.html>.
- Baratin, A., George, T., Laurent, C., Devon Hjelm, R., La joie, G., Vincent, P., and Lacoste-Julien, S. Implicit regularization via neural feature alignment. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2269–2277. PMLR, 13–15 Apr 2021. URL <https://proceedings.mlr.press/v130/baratin21a.html>.
- Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.
- Borup, K., Phoo, C. P., and Hariharan, B. Distilling from similar tasks for transfer learning on a budget. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11397–11407, 2023. doi: 10.1109/ICCV51070.2023.01050.
- Brown, D., Shapiro, M. R., Bittner, A., Warley, J., and Kvinge, H. Wild comparisons: A study of how representation similarity changes when input data is drawn from a shifted distribution. In *ICLR 2024 Workshop on Representational Alignment*, 2024. URL <https://openreview.net/forum?id=xapkeSwqf0>.
- Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., and Kim, S. Coyo-700m: Image-text pair dataset. <https://gitHub.com/kakaobrain/coyo-dataset>, 2022.
- Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020a. URL <https://proceedings.mlr.press/v119/chen20j.html>.
- Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 22243–22255. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fcbc95ccdd551da181207c0c1400c655-Paper.pdf.
- Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv e-prints*, 2020c. URL <https://arxiv.org/abs/2003.04297>.
- Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A. V., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B. K., Ruiz, C. R., Steiner, A. P., Angelova, A., Zhai, X., Houlsby, N., and Soricut, R. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=mWVoBz4W0u>.
- Cherti, M. and Beaumont, R. LAION-AI/CLIP_benchmark: Clip-like model evaluation, 2022. URL https://gitHub.com/LAION-AI/CLIP_benchmark.
- Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., and Konkle, T. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383, October 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-53147-y. URL <https://www.nature.com/article/10.1038/s41467-024-53147-y>.

- icles/s41467-024-53147-y. Publisher: Nature Publishing Group.
- Csiszik, A., Krosi-Szab, P., Matsangosz, A., Papp, G., and Varga, D. Similarity and matching of neural network representations. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5656–5668. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/2cb274e6ce940f47beb8011d8ecb1462-Paper.pdf.
- Cui, T., Kumar, Y., Marttinen, P., and Kaski, S. Deconfounded representation similarity for comparison of neural networks. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 19138–19151. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/79cbf4f96c2bcc67267421154da689dd-Paper-Conference.pdf.
- Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., Jenatton, R., Beyer, L., Tschannen, M., Arnab, A., Wang, X., Riquelme Ruiz, C., Minderer, M., Puigcerver, J., Evci, U., Kumar, M., Steenkiste, S. V., Elsayed, G. F., Mahendran, A., Yu, F., Oliver, A., Huot, F., Bastings, J., Collier, M., Gritsenko, A. A., Birodkar, V., Vasconcelos, C. N., Tay, Y., Mensink, T., Kolesnikov, A., Pavetic, F., Tran, D., Kipf, T., Lucic, M., Zhai, X., Keysers, D., Harmsen, J. J., and Houlsby, N. Scaling vision transformers to 22 billion parameters. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7480–7512. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/dehghani23a.html>.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- Ding, F., Denain, J.-S., and Steinhardt, J. Grounding representation similarity through statistical testing. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1556–1568. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0c0bf917c7942b5a08df71f9da626f97-Paper.pdf.
- Ding, Y., Tang, J., and Guo, F. Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing*, 325: 211–224, 2019. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2018.10.028>. URL <https://www.sciencedirect.com/science/article/pii/S0925231218312165>.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.
- Duong, L., Zhou, J., Nassar, J., Berman, J., Olieslagers, J., and Williams, A. H. Representational dissimilarity metric spaces for stochastic neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=xjb563TH-GH>.
- Dwivedi, K. and Roig, G. Representation similarity analysis for efficient task taxonomy & transfer learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12379–12388, 2019. doi: 10.1109/CVPR.2019.01267.
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bygh9j09KX>.
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020a. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL <https://doi.org/10.1038/s42256-020-00257-z>.
- Geirhos, R., Meding, K., and Wichmann, F. A. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13890–13902. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9f6992966d4c363ea0162a056cb45fe5-Paper.pdf.

- Geirhos, R., Narayananappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., and Brendel, W. Partial success in closing the gap between human and machine vision. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23885–23899. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c8877cff22082a16395a57e97232bb6f-Paper.pdf.
- Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=S1v4N210->.
- Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In Jain, S., Simon, H. U., and Tomita, E. (eds.), *Algorithmic Learning Theory*, pp. 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
- Hacohen, G., Choshen, L., and Weinshall, D. Let’s agree to agree: Neural networks share classification order on real datasets. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3950–3960. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/hacohen20a.html>.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.
- Hermann, K. and Lampinen, A. What shapes feature representations? exploring datasets, architectures, and training. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9995–10006. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/71e9c6620d381d60196ebe694840aaaa-Paper.pdf.
- Hernandez, A., Dangovski, R., Lu, P. Y., and Soljacic, M. Model stitching: Looking for functional similarity between representations. In *SVRHM 2022 Workshop @ NeurIPS*, 2022. URL <https://openreview.net/forum?id=7bHLC05FQdB>.
- Hotelling, H. *Relations Between Two Sets of Variates*, pp. 162–190. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_14. URL https://doi.org/10.1007/978-1-4612-4380-9_14.
- Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely Connected Convolutional Networks . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.243. URL <https://doi.ieee.org/10.1109/CVPR.2017.243>.
- Huh, M., Cheung, B., Wang, T., and Isola, P. Position: The platonic representation hypothesis. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/huh24a.html>.
- KakaoBrain. Align-base model, 2023. URL <https://huggingface.co/kakaobrain/align-base>. Accessed: 01.09.2024.
- Klabunde, M., Amor, M. B., Granitzer, M., and Lemmerich, F. Towards measuring representational similarity of large language models. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023. URL https://openreview.net/forum?id=gG5U_GgUzDR.
- Klabunde, M., Schumacher, T., Strohmaier, M., and Lemmerich, F. Similarity of neural network models: A survey of functional and representational measures. *ACM Comput. Surv.*, 57(9), May 2025. ISSN 0360-0300. doi: 10.1145/3728458. URL <https://doi.org/10.1145/3728458>.
- Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019a. URL <https://proceedings.mlr.press/v97/kornblith19a.html>.

- Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, 2019b. doi: 10.1109/CVPR.2019.00277.
- Kornblith, S., Chen, T., Lee, H., and Norouzi, M. Why do better loss functions lead to less transferable features? In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28648–28662. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f0bf4a2da952528910047c31b6c2e951-Paper.pdf.
- Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL <https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008>.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012a. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012b. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- Lake, B. M. and Baroni, M. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.
- Lampinen, A. K., Chan, S. C., and Hermann, K. Learned feature representations are biased by complexity, learning order, position, and more. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=aY2nsgE97a>.
- Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, Mar 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08987-4. URL <https://doi.org/10.1038/s41467-019-08987-4>.
- Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=B1EA-M-0Z>.
- Li, Y., Yosinski, J., Clune, J., Lipson, H., and Hopcroft, J. Convergent learning: Do different neural networks learn the same representations? In Storcheus, D., Rosamizadeh, A., and Kumar, S. (eds.), *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 of *Proceedings of Machine Learning Research*, pp. 196–212, Montreal, Canada, 11 Dec 2015. PMLR. URL <https://proceedings.mlr.press/v44/li15convergent.html>.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, 2022. doi: 10.1109/CVPR52688.2022.01167.
- Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=Skq89Scxx>.
- Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Maniparambil, M., Akshulakov, R., Dahou Djilali, Y. A., Seddik, M. E. A., Narayan, S., Mangalam, K., and O’Connor, N. E. Do vision and language encoders represent the world similarly? In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14334–14343, 2024. doi: 10.1109/CVPR52733.2024.01359.
- Matthews, A., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=H1-nGgWC->.
- Mehrer, J., Spoerer, C. J., Kriegeskorte, N., and Kietzmann, T. C. Individual differences among deep neural network models. *Nature Communications*, 11(1):5725, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19632-w.

- URL <https://doi.org/10.1038/s41467-020-19632-w>.
- Merullo, J., Castricato, L., Eickhoff, C., and Pavlick, E. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=8tYRqb05pVn>.
- Misra, I. and van der Maaten, L. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6706–6716, 2020. doi: 10.1109/CVPR4260.0.2020.00674.
- Morcos, A., Raghu, M., and Bengio, S. Insights on representational similarity in neural networks with canonical correlation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a7a3d70c6d17a73140918996d03c014f-Paper.pdf.
- Muttenhaler, L. and Hebart, M. N. THINGSvision: A python toolbox for streamlining the extraction of activations from deep neural networks. *Frontiers in Neuroinformatics*, 15:45, 2021. ISSN 1662-5196. doi: 10.3389/fninf.2021.679838. URL <https://www.frontiersin.org/article/10.3389/fninf.2021.679838>.
- Muttenhaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., and Kornblith, S. Human alignment of neural network representations. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=ReDQ1OUQR0X>.
- Muttenhaler, L., Greff, K., Born, F., Spitzer, B., Kornblith, S., Mozer, M. C., Müller, K.-R., Unterthiner, T., and Lampinen, A. K. Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*, 2024.
- Nguyen, T., Raghu, M., and Kornblith, S. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=KJNcAkY8tY4>.
- Noroosi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision – ECCV 2016*, pp. 69–84. Springer International Publishing, 2016.
- Oquab, M., Dariseti, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. DINoV2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=a68SUT6zFt>.
- Princeton University. About WordNet. <https://wordnet.princeton.edu/>, 2010.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.
- Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. Do vision transformers see like convolutional neural networks? In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/652cf38361a209088302ba2b8b7f51e0-Paper.pdf.
- Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik, L. Imagenet-21k pretraining for the masses. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/98f13708210194c475687be6106a3b84-Paper-round1.pdf.
- Saha, A., Bialkowski, A. N., and Khalifa, S. Distilling representational similarity using centered kernel align-

- ment (cka). In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. URL <https://bmvc2022.mpi-inf.mpg.de/0535.pdf>.
- Santurkar, S., Tsipras, D., and Madry, A. BREEDS: benchmarks for subpopulation shift. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=mQPBMvyAuk>.
- Schuhmann, C., Kaczmarczyk, R., Komatsuzaki, A., Katta, A., Vencu, R., Beaumont, R., Jitsev, J., Coombes, T., and Mullis, C. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021.
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crownson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. Laion-5b: An open large-scale dataset for training next generation image-text models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25278–25294. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debf3b3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1409.1556>.
- Steiner, A. P., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL <https://openreview.net/forum?id=4nPswr1Kcp>.
- Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., Tenenbaum, J. B., Collins, K. M., Hermann, K. L., Oktar, K., Greff, K., Hebart, M. N., Jacoby, N., Zhang, Q., Marjeh, R., Geirhos, R., Chen, S., Kornblith, S., Rane, S., Konkle, T., O’Connell, T. P., Unterthiner, T., Lampinen, A. K., Müller, K.-R., Toneva, M., and Griffiths, T. L. Getting aligned on representational alignment, 2023. URL <https://arxiv.org/abs/2310.13018>.
- Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. Evaclip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.
- Tan, M. and Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/tan19a.html>.
- Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., and Summerfield, C. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024. doi: 10.1126/science.adq2852. URL <https://www.science.org/doi/abs/10.1126/science.adq2852>.
- Tian, Y., Krishnan, D., and Isola, P. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SkgpBJrtvS>.
- Tian, Y., Fan, L., Chen, K., Katabi, D., Krishnan, D., and Isola, P. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15887–15898, June 2024.
- Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., and Dosovitskiy, A. Mlp-mixer: An all-mlp architecture for vision. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24261–24272. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/cba0a4ee5cccd02fda0fe3f9a3e7b89fe-Paper.pdf.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. Training data-efficient image transformers & distillation through attention. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/touvron21a.html>.

- <https://proceedings.mlr.press/v139/touvron21a.html>.
- Williams, A. H., Kunz, E., Kornblith, S., and Linderman, S. Generalized shape metrics on neural representations. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/252a3dbaeb32e7690242ad3b556e626b-Paper.pdf.
- Xu, Y. and Vaziri-Pashkam, M. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, 12(1):2065, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22244-7. URL <https://doi.org/10.1038/s41467-021-22244-7>.
- Yang, G. and Hu, E. J. Tensor programs iv: Feature learning in infinite-width neural networks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/yang21c.html>.
- Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/zbontar21a.html>.
- Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., and Houlsby, N. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2020. URL <https://arxiv.org/abs/1910.04867>.
- Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.
- Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.
- Zong, M., Qiu, Z., Ma, X., Yang, K., Liu, C., Hou, J., Yi, S., and Ouyang, W. Better teacher better student: Dynamic prior knowledge for knowledge distillation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=M0_sUuEyHs.

Appendix

A. Models and Datasets

We considered 23 downstream datasets (Tab. 1) and 64 pretrained vision models (Tab. 2) in our analysis. For each dataset, we selected the training split, potentially subsetted for large datasets as described in Appx. C, to compute the model representational similarities, while we use the validation or test split to compute downstream performance.

Table 1: Datasets of clip-benchmark and breeds and their domains according to (Zhai et al., 2020) with additional separation of natural images into multi- and single-domain.

Natural (multi-domain)	Natural (single-domain)	Specialized	Structured
Caltech101	Cars	Diabetic Retinopathy	FER2013
CIFAR-10	FGVC Aircraft	EuroSAT	Dmlab
CIFAR-100	Flowers	PatchCamelyon (PCam)	DTD
Country211	GTSRB	RESISC45	
ImageNet-1k	Pets		
STL-10	SVHN		
Pascal VOC 2007			
Entity-30 (Breeds)			
Living-17 (Breeds)			
NonLiving-26 (Breeds)			

Models. The models were chosen to represent a diverse set of training datasets, training objectives, model sizes, and architectures. Training objectives include standard supervised learning (Sup), self-supervised learning (SSL), and image-text alignment (Img-Txt). The models have been trained on datasets spanning four classes of increasing scale and semantic diversity. The most focused class consists of ImageNet-1k (IN1k; (Deng et al., 2009)) with 1,000 object categories, followed by models trained on ImageNet-21k (IN21k; (Ridnik et al., 2021)) or a combination of IN21k and IN1k, expanding to 21,000 categories. The "Large" class encompasses COYO-700M (Byeon et al., 2022), LAION400M (Schuhmann et al., 2021), LVD-142M (Oquab et al., 2024), and WIT-400M (Radford et al., 2021), which move beyond curated object categories to include broader visual concepts and text descriptions. Finally, the "XLarge" class includes LAION2B (English subset of LAION5B (Schuhmann et al., 2022)), Merged2B (merged version of LAION2B and COYO-700M; (Sun et al., 2023)), and WebLI (Chen et al., 2023), which contain billions of image-text pairs covering an extremely wide range of visual and semantic content. In terms of architecture, the models employ a range of designs, including Convolutional Neural Networks (CNNs) such as ResNet, EfficientNet, ConvNeXt, and VGG, as well as Transformer-based models like ViT and Swin-Transformer. Further, we divide the models into four different size categories, ranging from small, with around 12 million parameters, to xlarge, with over 1.4 billion parameters. Tab. 3 summarizes the number of models in each category, and Fig. 9 shows the distribution of models across pairs of categories, highlighting the relationships between different model attributes.

Table 3: Number of models in each attribute category.

Category	Nr. models	Category	Nr. models
Training objective		Training data	
Image-Text (Img-Txt)	14	ImageNet-1k (IN1k)	37
Self-Supervised (SSL)	20	ImageNet-21k (IN21k)	9
Supervised (Sup)	30	Large	11
		XLarge	7
Architecture Class		Model Size	
Convolutional (CNN)	24	small < 100M parameter	32
Transformer (TX)	40	medium < 200M parameter	14
		large < 400M parameter	8
		xlarge > 400M parameter	10

Preprocessing. To extract the latent representations for each model-dataset combination, we used the Python package

thingsvision (Muttenthaler & Hebart, 2021). Images were resized to 256px and center-cropped to 224px before applying the model-specific normalizations from the pretraining. Tab. 2 specifies exactly which layers were used for each model. For ViT models, we used the CLS token representations after the final LayerNorm module. We applied L2-normalization to ensure all feature vectors are of unit length.

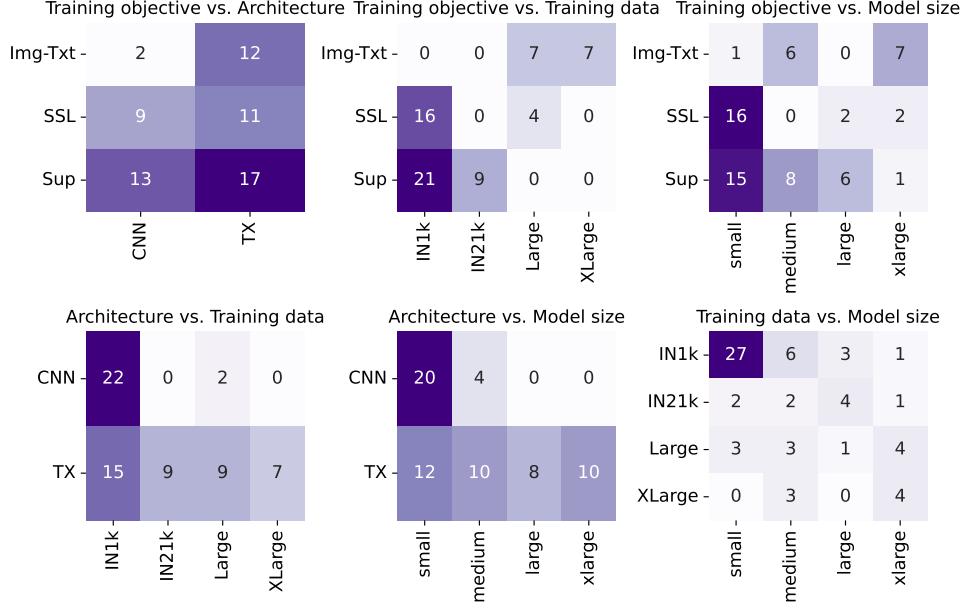


Figure 9: Distribution of models across paired categories. Each heatmap shows the count of vision models for different combinations of model categories: training objective (Sup: supervised, SSL: self-supervised learning, Img-Txt: image-text), architecture (CNN: convolutional neural networks, TX: transformers), training data (IN1k: ImageNet-1k, IN21k: ImageNet-21k, Large/XLarge: larger datasets), and model size (small to xlarge)

B. Downstream task performance

We evaluated each model’s downstream task performance on every dataset by calculating the top-1 accuracy of a linear probe trained on features extracted from the corresponding training set. Each model was evaluated with 3 different random seeds on each dataset, and the mean top-1 accuracy across seeds was used. Hyperparameter selection was performed with a validation set of 20% of the training set and the remaining training data for optimization. We followed the binary search procedure described in (Radford et al., 2021) and searched for the optimal weight decay parameter λ in the interval between 10^{-6} and 10^2 in 96 logarithmically spaced steps. This was done for all learning rates $\eta \in \{10^i\}_{i=1}^4$. After hyperparameter selection, the linear probe was retrained on the full training set and evaluated on the respective test set (validation set for Imagenet-1k). All linear probes are trained for 20 epochs, using the AdamW optimizer (Loshchilov & Hutter, 2019) and a cosine schedule for learning rate decay (Loshchilov & Hutter, 2017).

C. CKA sensitivity to the number of samples in dataset

Our analysis of model similarities requires reliable CKA measurements while managing computational constraints. To identify the minimum number of samples needed for stable CKA values, we analyzed ImageNet-1k subsets with varying samples per class ($k \in \{1, 5, 10, 20, 30, 40\}$). We generated pairwise similarity matrices (Fig. 10) and assessed stability by comparing the absolute differences between consecutive sample sizes (Fig. 11). The results show that most CKA variants converged with 10,000 samples ($k=10$), with only CKA RBF ($\sigma = 0.2$) requiring more samples (30,000, $k=30$) for stability.

To further assess CKA stability across different samples, we conducted bootstrapping experiments using six representative models (OpenCLIP ViT-L, OpenCLIP RN50, ResNet-50, ViT-L, DINoV2 ViT-L, and SimCLR RN50—the same anchor models described in Appx. G). For CKA linear, we performed 500 bootstrap iterations with 10,000 images per subset. For CKA RBF ($\sigma = 0.2$), we performed 500 bootstrap iterations with 30,000 images per subset. For each bootstrap sample, we

extracted features from all six models and computed CKA values between each model pair.

Fig. 12(a) confirms the stability of CKA linear values across all model pairs, while for CKA RBF ($\sigma = 0.2$), Fig. 12(b) reveals that even when using the larger 30,000 samples per subset, we still observe slightly higher variability compared to CKA linear.

These findings indicate that local similarity measures (RBF kernel) are more sensitive to specific stimuli than global similarity measures (linear kernel). This aligns with theoretical expectations, as stimulus-specific fine-grained details have a greater influence on local similarity measurements.

Based on these findings, we applied stratified subsampling to limit all datasets to 10,000 or 30,000 samples for CKA linear and CKA RBF ($\sigma = 0.2$), respectively.

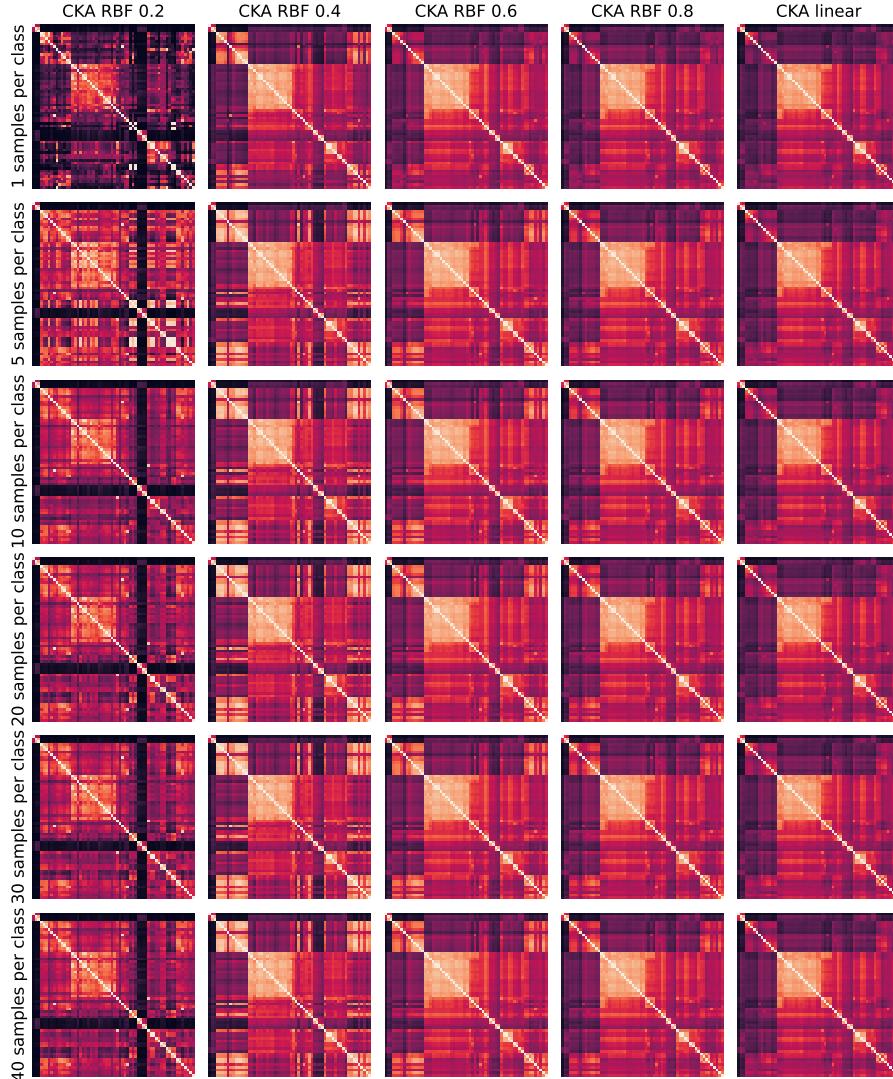


Figure 10: Pairwise model representational similarity matrices using different dataset sizes (1-40 samples per ImageNet-1k class) and metrics (CKA RBF 0.2-0.8 to linear).

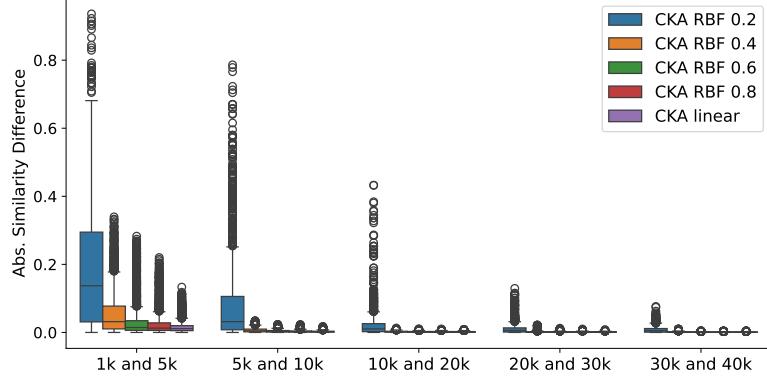


Figure 11: Distribution of absolute differences of pairwise model representational similarities for each CKA variant measured on subsets of ImageNet-1k with consecutive sample sizes per class ($1 \rightarrow 5, \dots, 30 \rightarrow 40$). Most CKA variants converge at $k = 10$ samples per class, with CKA RBF ($\sigma = 0.2$) requiring larger sample sizes for stability.

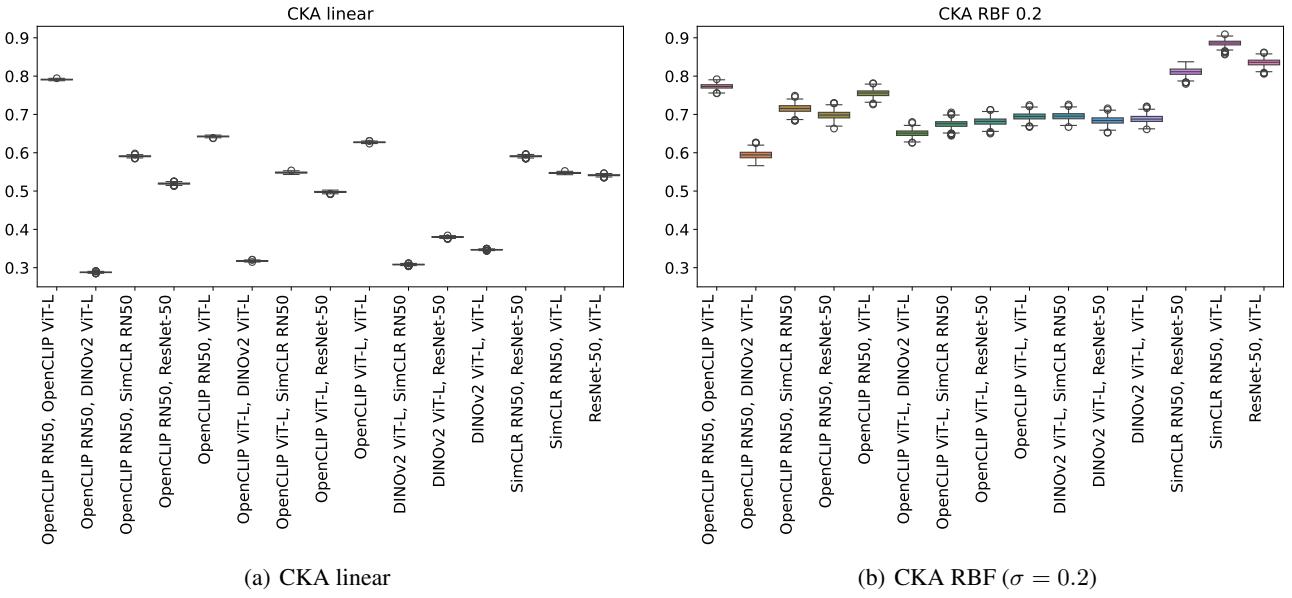


Figure 12: Distribution of CKA values between model representations across 500 randomly sampled ImageNet-1k subsets for different model pairs. (a) Linear CKA computed on 10k samples per subset. (b) RBF CKA ($\sigma = 0.2$) computed on 30k samples per subset.

D. Similarity matrices for different datasets

In § 4.2, we showed that similarity measurements are not directly transferable across datasets due to their high variability. Fig. 2 displays representational similarity matrices computed using CKA linear across three selected datasets, along with the mean and standard deviation matrices across all evaluation datasets. In Fig. 13, we show additional similarity matrices for three representative datasets from each dataset category (see Appx. A for details). The yellow-boxed models, containing image-text models, are most similar on natural image datasets. In contrast, the white-boxed models, containing self-supervised models, show the highest similarity on specialized EuroSAT and PCAM datasets. The cyan-boxed models, containing supervised models, are most dataset dependent, showing high similarity on Pets but low similarity on ImageNet-1k.

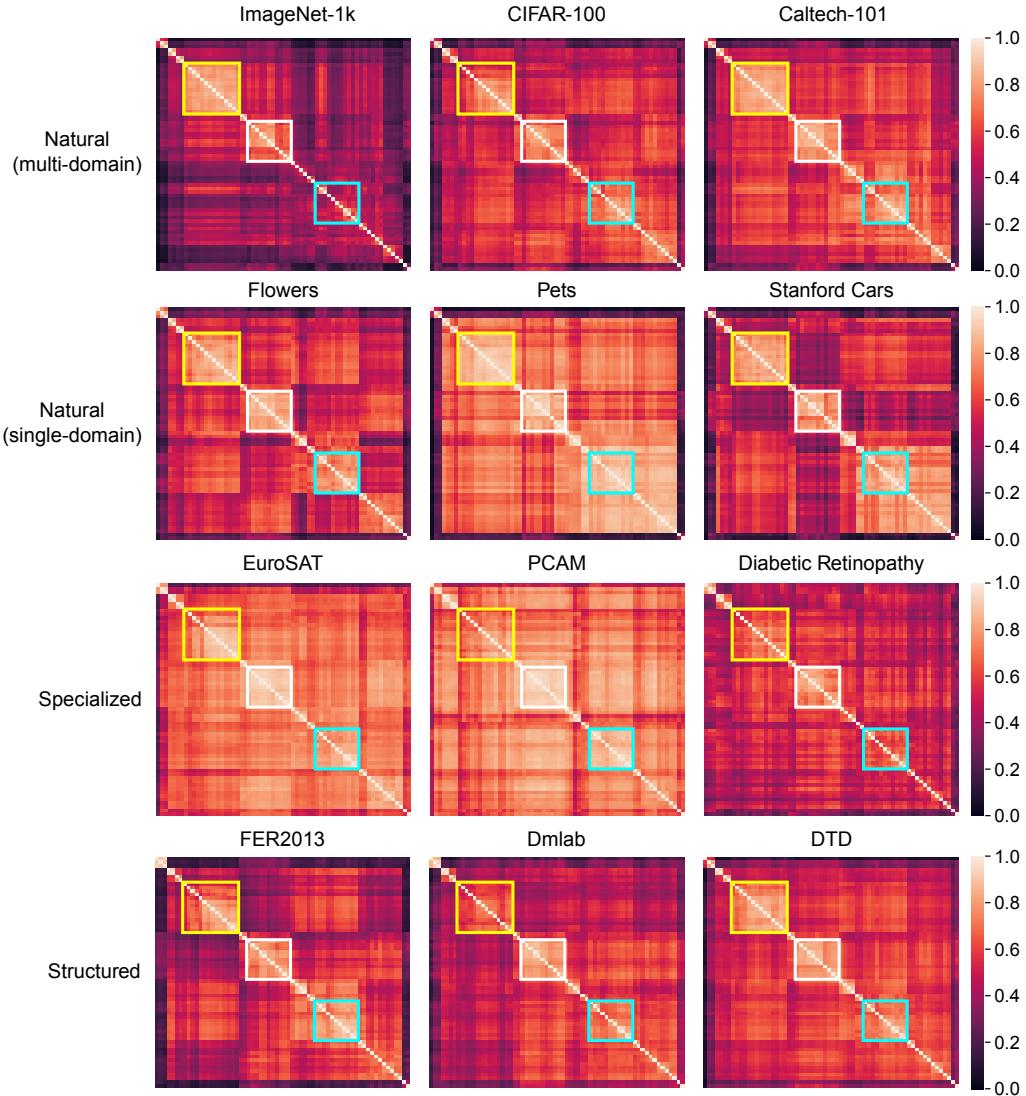


Figure 13: Representational similarity of model pairs, calculated with linear CKA. For each dataset category, three evaluation datasets are displayed.

E. Variation of CKA values near the upper bound

The CKA similarity measure is bounded between 0 and 1, with values close to 1 indicating highly similar representations. However, this upper bound might lead to saturation effects. To better differentiate between highly similar model representations, we apply two nonlinear transformations: \arccos and \tan . Both transformations expand the range of high CKA

values, with tan slightly reducing variation for low similarity values. These transformations amplify the inverse U-shape relationship noted in Section 4.2 (Fig. 14), suggesting that the relationship between mean and variance of similarities is robust across different scales.

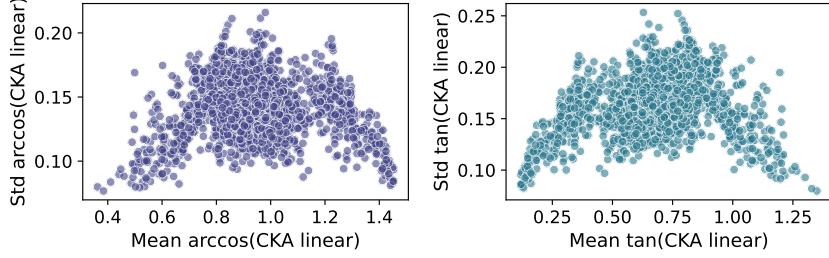


Figure 14: Mean versus standard deviation of CKA values after nonlinear transformations. **Left:** arccos transformation inverts the scale, mapping high similarities to low values and vice versa. **Right:** tan transformation. Both transformations make the inverse U-shape relationship more pronounced.

F. Influence of similarity metric on similarity consistency

In our main analysis, we used CKA linear as a global similarity metric. Fig. 3 shows strong positive correlations between mean representational similarities across datasets when measured with different metrics. Here, we investigate whether this stability extends to similarity consistency distributions when using alternative metrics.

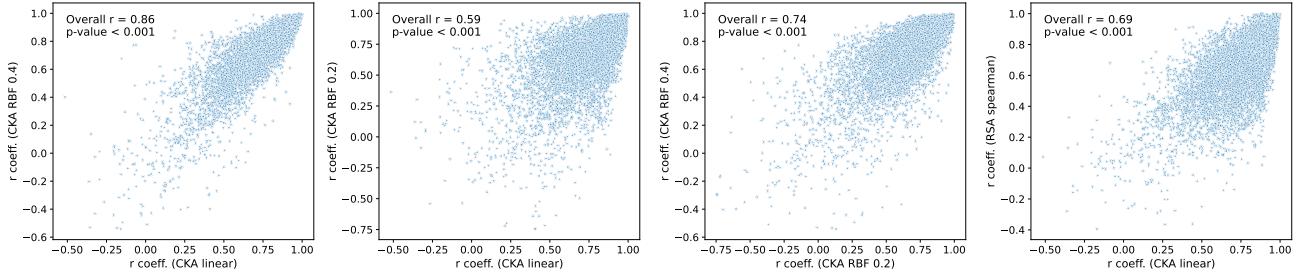


Figure 15: Relationship between similarity consistency values measured with different similarity metrics across all model sets and dataset pairs. From left to right: CKA linear vs. CKA RBF ($\sigma = 0.4$), CKA linear vs. CKA RBF ($\sigma = 0.2$), CKA RBF ($\sigma = 0.2$) vs. CKA RBF ($\sigma = 0.4$), and CKA linear vs. RSA Spearman. Correlations are measured using Pearson correlation. Each point represents the similarity consistency value for a pair of datasets and model sets.

Local similarity metrics When examining local similarity structures, we leverage the flexibility of CKA with an RBF kernel. The kernel bandwidth parameter σ controls sensitivity to local structure, allowing us to capture different aspects of representational similarity. Notably, Fig. 10 reveals that pattern differences between CKA linear and CKA RBF ($\sigma = 0.2$) are substantially more pronounced than those between CKA linear and CKA RBF ($\sigma = 0.4$). This suggests that the latter partially captures global similarity structures rather than purely local ones. Indeed, the correlation between similarity consistency values measured with CKA linear and CKA RBF ($\sigma = 0.4$) is strong (Fig. 15, leftmost panel), while it is substantially lower between CKA linear vs. CKA RBF ($\sigma = 0.2$) (Fig. 15, second panel). The correlation between similarity consistency values measured with the two different sigmas ($\sigma = 0.2$ and $\sigma = 0.4$) (Fig. 15, third panel) is in between, strengthening the hypothesis that CKA RBF ($\sigma = 0.4$) partially captures global structure. Therefore, we use CKA RBF ($\sigma = 0.2$) as our local similarity metric.

We observe the same overall pattern as in Fig. 6 of the main text for the CKA RBF ($\sigma = 0.2$) similarity metric: the objective is a main influencing factor for similarity consistency, while network architecture and model size are less important (Fig. 16). However, the distributions of consistency values reveal two interesting observations. For local similarity, supervised model pairs (Φ_{Sup} , Φ_{Sup}) are more consistent than the image-text model pairs ($\Phi_{\text{Img-Txt}}$, $\Phi_{\text{Img-Txt}}$), which are better at representing global structure. In addition, models trained on IN21k (Φ_{IN21k} , Φ_{IN21k}) are more consistent than models trained on IN1k

$(\Phi_{IN1k}, \Phi_{IN1k})$). Their ordering flipped compared to Fig. 6, where $(\Phi_{IN1k}, \Phi_{IN1k})$ model pairs are more consistent in their global structure. IN21k contains substantially more classes (21.843) and a higher percentage of the classes are leaf nodes in the WordNet (Princeton University, 2010) tree (76.71% for IN21k vs. 65% for IN1k), representing more fine-grained entities. The representation must contain more fine-grained details to distinguish these classes, dominating local similarities.

Other global similarity metrics In Fig. 17, the correlation between the two global measures (CKA linear and RSA Spearman) is moderate. Even with this reduced correlation strength, similarity consistency measured using RSA Spearman exhibits the same patterns as CKA linear, particularly regarding the model category we found to be most influential – the training objective (see Fig. 17).

The analysis presented in this section suggests that our observations on the influence of the training objective on similarity consistency hold for multiple metrics.

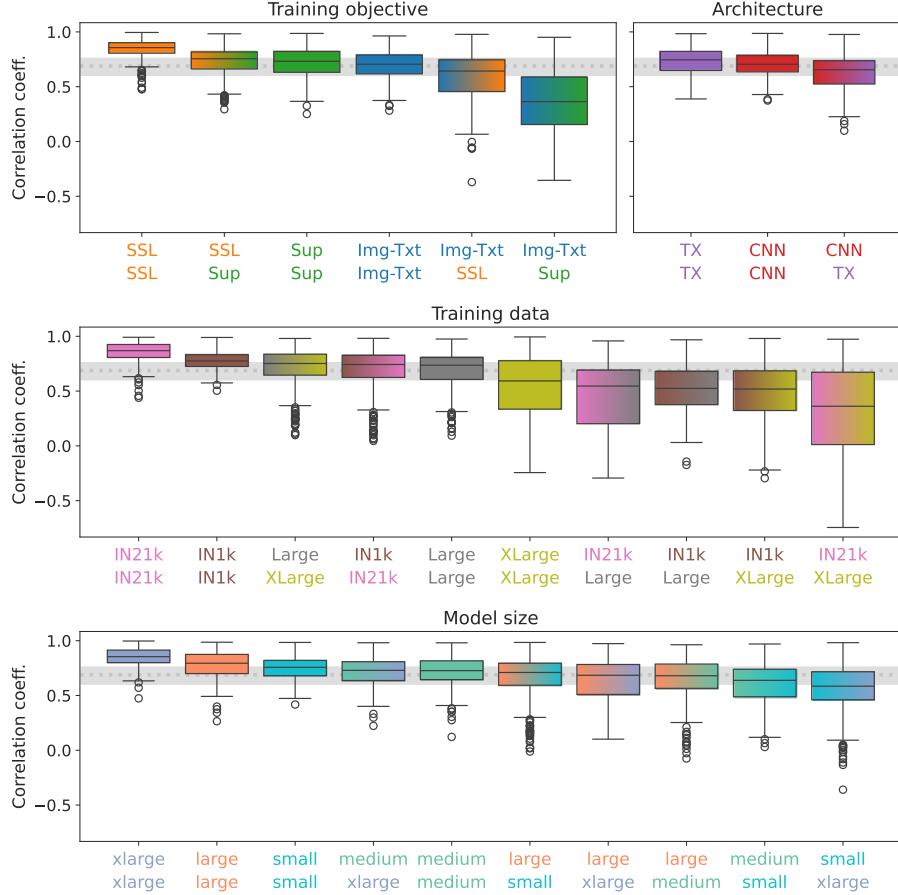


Figure 16: Distribution of similarity consistencies for each model set pair based on CKA RBF ($\sigma = 0.2$) representational similarities. The boxes are sorted in decreasing median correlation. The dotted line indicates the overall median, while the gray area spans the 25th to 75th percentiles of correlations across all model pairs.

G. Validating training objective effects through controlled anchor model analysis

To validate our findings regarding the influence of different model categories (§4.5), we conducted a controlled analysis using six carefully selected anchor models: OpenCLIP RN50, OpenCLIP ViT-L, SimCLR RN50, DINOv2 ViT-L, ResNet-50, and ViT-L. These models were chosen to create systematic variations across training objectives (image-text, self-supervised, and supervised learning), while controlling for architecture (convolutional ResNet-50 vs transformer ViT-L) and model size. This controlled setup helps verify whether the patterns observed in our main analysis persist when comparing individual representative models to broader model sets. For each anchor model, we define a single-element model set (e.g.,

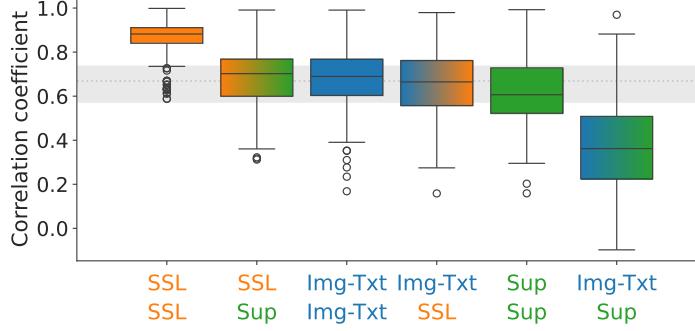


Figure 17: Distribution of similarity consistency values for each model set pair within the training objective category, based on RSA Spearman representational similarities

$\Phi_{\text{OpenCLIP RN50}}$) and compute its similarity consistency with different model groups (e.g., $R(\Phi_{\text{OpenCLIP RN50}}, \Phi_{\text{SSL}})$.

Fig. 18 shows the correlation coefficient distributions across dataset pairs, focusing on training objectives as they emerged as the most influential factor for similarity consistency. The results support our main findings while providing additional insights. Self-supervised models (SimCLR RN50, DINOv2 ViT-L) consistently show the highest median correlations and lowest variations when compared with any anchor model, regardless of the anchor's training objective. Surprisingly, for global similarity, even image-text (OpenCLIP RN50, OpenCLIP ViT-L) and supervised (ResNet-50, ViT-L) anchor models correlate more strongly with Φ_{SSL} than with their own categories ($\Phi_{\text{Img-Txt}}$ and Φ_{Sup} , respectively). However, the image-text anchors show high variation of similarity consistency for local similarity measures and all training objectives. We also observe a notable pattern of weak correlations between image-text anchors and supervised models (Φ_{Sup}) and vice versa. These patterns remain remarkably consistent across different architectures and model sizes, suggesting that training objectives, rather than architectural choices or model sizes, primarily drive similarity relationships across datasets.

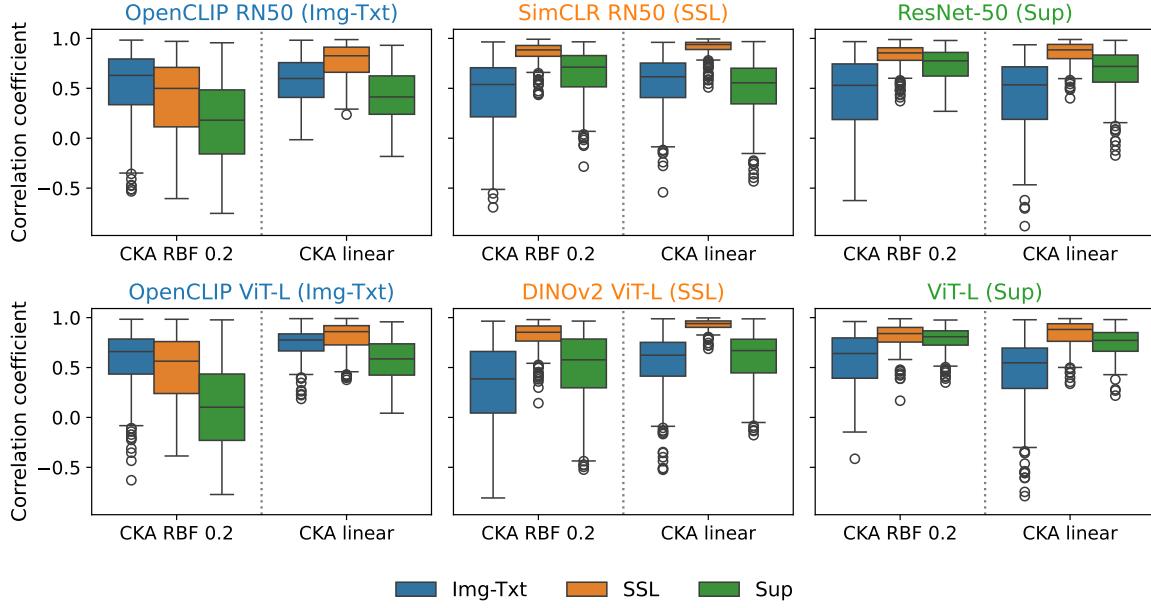


Figure 18: Distribution of Pearson correlation coefficients between anchor models (OpenCLIP RN50, OpenCLIP ViT-L, SimCLR RN50, DINOv2 ViT-L, ResNet-50, ViT-L) and models trained with different objectives. Each box shows correlations across all dataset pairs between one anchor model and models from specific sets (e.g., Φ_{SSL} , Φ_{Sup}). Results for CKA RBF ($\sigma = 0.2$) and CKA linear are separated by dotted lines. Subplot titles are colored according to the anchor model's subcategory.

H. Effect of evaluation data on similarity consistency

In the main analysis, we examined similarity consistency across model categories for all evaluation datasets (§4.5) and across individual datasets for all models (§4.6). Here, we investigate whether model category effects differ between natural and non-natural (specialized and structured) datasets, as non-natural datasets represent out-of-distribution data for many models (Fig. 19).

The primary finding from §4.5 holds: training objective remains the dominant factor influencing similarity consistency regardless of dataset type. However, similarity consistency between image-text and supervised models ($\Phi_{\text{Img-Txt}}$, Φ_{Sup}) shows lower medians and higher variance on natural datasets compared to non-natural ones. For natural datasets, supervised models might encode features tightly coupled to class-label supervision, which diverge from the semantically richer, language-aligned representations of image-text models. Conversely, on non-natural datasets where both model types operate further from their training distributions, this representational gap narrows as both might rely on more generalized features. Complementary, the analysis—examining how training dataset composition affects these patterns—is presented in the next section.

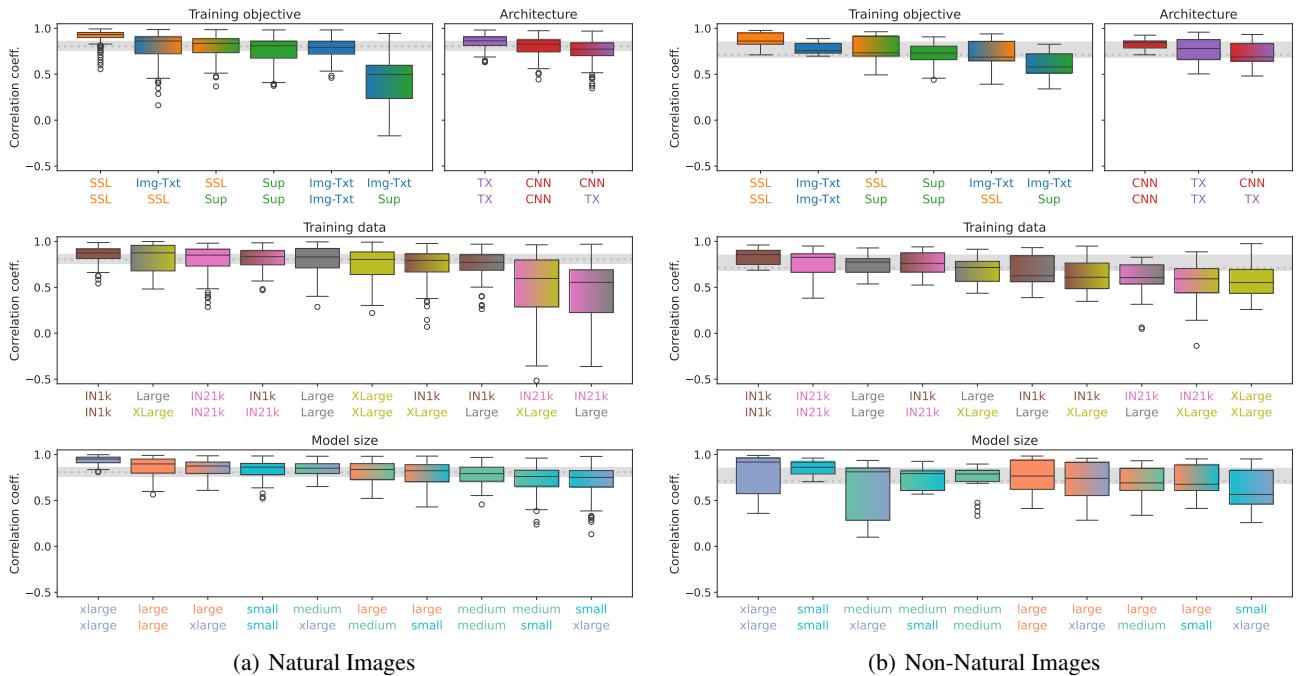


Figure 19: Distribution of similarity consistencies for natural (single- and multi-domain) and non-natural (specialized and structured) datasets. Representational similarities for each model set pair are computed with CKA linear.

I. Effect of training data domain on similarity consistency

In our main analysis, we restricted our focus to general-purpose vision models trained on large-scale datasets with diverse semantic classes (§ 4.1). This focus was motivated by the Platonic representation hypothesis, which suggests that as the semantic diversity of training data approaches the complexity of reality itself, representations should increasingly converge (Huh et al., 2024). However, Conwell et al. (2024) demonstrated that training data is a key driver of representational similarity between models, potentially more significant than architecture, training objective, or choice of similarity metric. In this section, we examine how this observation manifests in terms of representational consistency.

To investigate the role of training data domain, we selected four architectures (AlexNet (Krizhevsky et al., 2012a), DenseNet161 (Huang et al., 2017), ResNet18 (He et al., 2016), and ResNet50 (He et al., 2016)) with available pre-trained weights for both the general-purpose ImageNet-1k and the scene-specific Places365 (Zhou et al., 2018) datasets. Models

were sourced from torchvision¹, using default ImageNet weights and Places365 weights from the official GitHub repository². This selection of models allowed us to isolate the training data effect on similarity consistency while fixing the other characteristics: training objective (Supervised), architecture type (CNN), and model size (small).

We evaluated similarity consistency across the 23 downstream datasets for all model set pairs: $(\Phi_{IN1k}, \Phi_{IN1k})$, $(\Phi_{IN1k}, \Phi_{Places365})$, $(\Phi_{Places365}, \Phi_{Places365})$, and (Φ_{all}, Φ_{all}) , where Φ_{all} contains all models.

Our analysis reveals two main observations. First, as shown in Fig. 20, Places365-trained model pairs demonstrate more consistent representational similarities compared to ImageNet-1k-trained pairs. We hypothesize that domain-specific training constrains the solution space more tightly, whereas the greater diversity of semantic classes in ImageNet-1k allows for more varied learned representations across models. This distinction is particularly interesting in light of the platonic representation hypothesis. While domain-specific models show higher consistency, this may reflect convergence to specialized solutions rather than to general representations of reality. In contrast, the lower consistency among ImageNet-1k-trained models likely reflects the challenge of learning representations that capture broad semantic diversity. This consistency might be even lower for models trained with supervised objectives since they can exploit different confounders and shortcuts when learning to classify the diverse set of classes.

Second, the distributions $R(\Phi_{IN1k}, \Phi_{IN1k})$, $R(\Phi_{IN1k}, \Phi_{Places365})$, and $R(\Phi_{all}, \Phi_{all})$ exhibit similar distribution spreads. Additionally, $R(\Phi_{all}, \Phi_{all})$ shows a distribution similar to $R(\Phi_{IN1k}, \Phi_{IN1k})$ in Fig. 6. This suggests that including domain-specific models alongside general-purpose ones might not fundamentally alter the representational similarity consistency patterns observed in our analysis.

Our analysis shows that the training data domain significantly influences representational similarity consistency, though mixing domain-specific and general-purpose models preserves the broad distribution patterns of correlation coefficients observed in ImageNet-1k models alone.

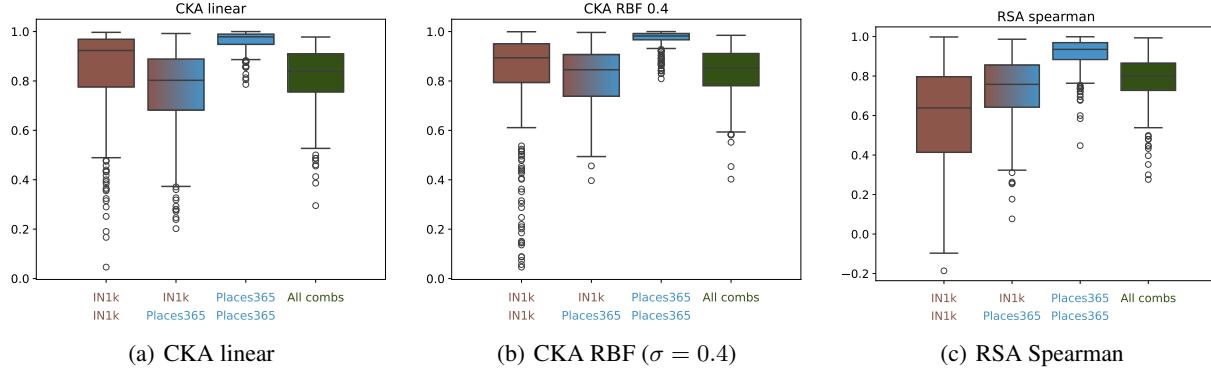


Figure 20: Distribution of Pearson correlation coefficients for model similarities computed with different similarity metrics across dataset pairs, i.e., $R(\Phi_{IN1k}, \Phi_{IN1k})$, $R(\Phi_{IN1k}, \Phi_{Places365})$, $R(\Phi_{Places365}, \Phi_{Places365})$, and $R(\Phi_{all}, \Phi_{all})$, with Φ_{IN1k} containing the four IN1k and $\Phi_{Places365}$ containing the four Places365 pre-trained models. Φ_{all} contains all eight models.

J. Distribution correlation coefficients downstream task performance vs. model similarity

In §4.7, we examined the correlations between representational similarity and classification performance gaps across three datasets from each of the four dataset categories: Natural (multi-domain), Natural (single-domain), Specialized, and Structured. We observed stronger negative correlations for the Natural (single-domain) datasets. Fig. 22 extends this analysis by illustrating the correlation coefficients for all evaluated datasets. We can confirm that natural (single-domain) datasets exhibit the strongest negative correlations overall on this larger set of datasets, with the SVHN dataset being the sole outlier.

Interestingly, ImageNet-1k exhibits the weakest negative correlation between performance gap and representational similarity.

¹<https://pytorch.org/vision/stable/index.html>

²<https://github.com/CSAILVision/places365>

We hypothesize that this is due in part to the high diversity of classes in ImageNet-1k, along with its rich set of confounders and contextual cues, which enable networks to approach the classification task in multiple ways. This may be exacerbated by the fact that some of the evaluated networks were pretrained on ImageNet-1k or ImageNet-21k, which could have encouraged the learning of dataset-specific features that perform well on ImageNet but fail to generalize to other datasets.

While our categorization of natural datasets into single-domain and multi-domain revealed an interesting difference in the relationship between downstream task performance and model similarity—especially between the ImageNet-1k and single-domain datasets like Flowers or Pets—the correlation coefficients for these categories still overlap. We hypothesize that our binary sub-categorization of Natural datasets is insufficient for capturing a clear distinction for two reasons: (1) Datasets lie on a spectrum between the two (hypothetical) extremes of encompassing all possible domains and encompassing exactly one. For instance, ImageNet-1k has a relatively rich set of domains, containing 1,000 diverse classes, whereas CIFAR-10 only contains 10 classes of vehicles and animals. (2) Datasets differ in the amount of contextual information (e.g., the object’s background) they offer. We assume that contextual information plays an important role in the number of viable features a dataset classification can be solved with. A large number of viable features means that multiple well-performing but dissimilar strategies may exist. Consequently, multi-domain datasets with limited contextual information may behave more like single-domain datasets. For example, the PASCAL VOC 2007 dataset, when used for single-label classification, makes use of tight bounding-box crops around the object to be classified, thereby reducing the available context. Similarly, in low-resolution datasets such as CIFAR-10 and CIFAR-100, contextual information may be reduced due to the loss of background information. While we have not quantified it, we found the Caltech-101 dataset to contain much cleaner object backgrounds than ImageNet-1k, also suggesting reduced contextual information.

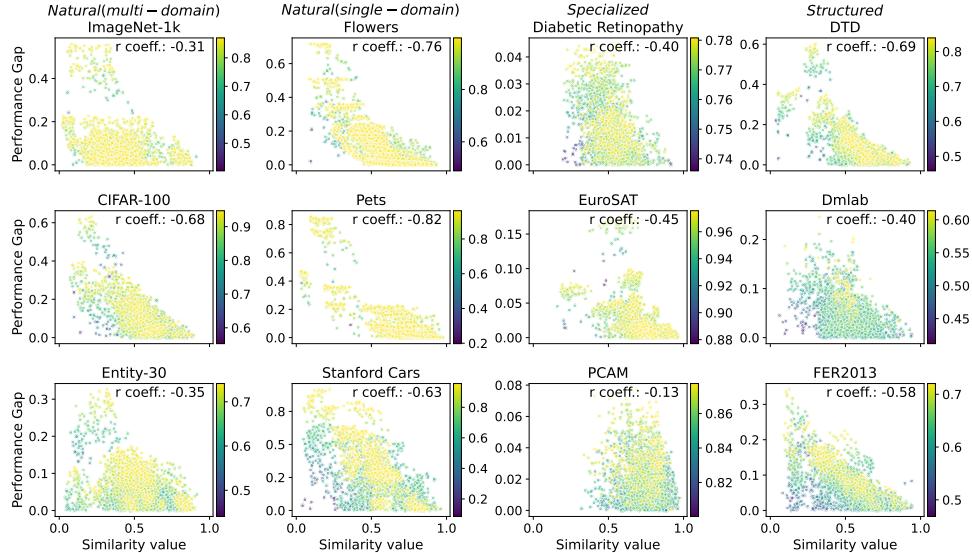


Figure 21: Model similarity (CKA linear) vs. absolute difference in downstream task performance (top-1 accuracy) for model pairs for three datasets per dataset category (natural multi-domain, natural single-domain, specialized, and structured). The color of each point indicates the downstream task accuracy of the better-performing model. Pearson correlation coefficients are shown in each subplot.

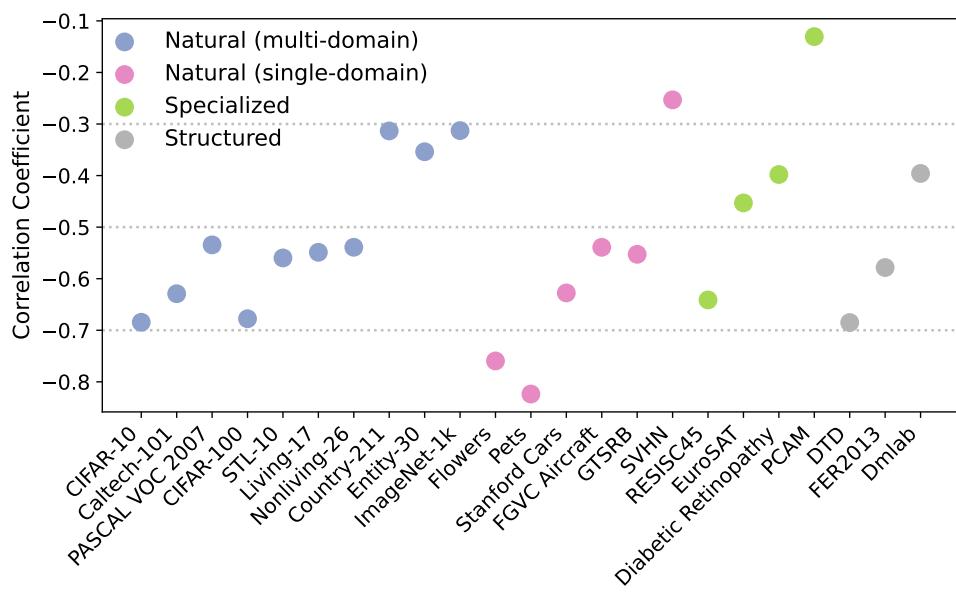


Figure 22: Pearson correlation coefficients between downstream task performance gap and model representation similarity for all 23 datasets grouped by dataset category.

Table 2: Pretrained neural networks that we considered in our analyses.

Model name	Source	Rep. Layer	Training objective	Training data	Training data class	Architecture	Architecture class	Model size	Model size class
Kakabrain-Align	KakaBrain (2023)	pooler	Img-Txt	COYO-700M	Large	EfficientNet	CNN	62.1M	small
OpenCLIP-EVA01-g-14-plus-merged2b-s11b-b114k	Sun et al. (2023)	visual	Img-Txt	Merged2B	XLarge	ViT	TX	1.4B	xlarge
OpenCLIP-EVA01-g-14-laion400m-s11b-b41k	Sun et al. (2023)	visual	Img-Txt	LAION400M	Large	ViT	TX	1.1B	xlarge
OpenCLIP-EVA02-B-16-merged2b-s8b-b131k	Sun et al. (2023)	visual	Img-Txt	Merged2B	XLarge	ViT	TX	149.7M	medium
OpenCLIP-EVA02-L-14-merged2b-s4b-b131k	Sun et al. (2023)	visual	Img-Txt	Merged2B	XLarge	ViT	TX	427.8M	xlarge
OpenCLIP-RN50-openai	Radford et al. (2021)	visual	Img-Txt	WIT-400M	Large	ResNet	CNN	102.0M	medium
OpenCLIP-ViT-B-16-SigLIP-webli	Zhai et al. (2023)	visual	Img-Txt	WebLi	XLarge	ViT	TX	203.2M	medium
OpenCLIP-ViT-B-16-laion2b-s34b-b88k	Radford et al. (2021)	visual	Img-Txt	LAION2B	XLarge	ViT	TX	149.6M	medium
OpenCLIP-ViT-B-16-laion400m-e32	Radford et al. (2021)	visual	Img-Txt	LAION400M	Large	ViT	TX	149.6M	medium
OpenCLIP-ViT-B-16-openai	Radford et al. (2021)	visual	Img-Txt	WIT-400M	Large	ViT	TX	149.6M	medium
OpenCLIP-ViT-L-14-laion2b-s32b-b82k	Radford et al. (2021)	visual	Img-Txt	LAION2B	XLarge	ViT	TX	427.6M	xlarge
OpenCLIP-ViT-L-14-laion400m-e32	Radford et al. (2021)	visual	Img-Txt	LAION400M	Large	ViT	TX	427.6M	xlarge
OpenCLIP-ViT-L-14-openai	Radford et al. (2021)	visual	Img-Txt	WIT-400M	Large	ViT	TX	427.6M	xlarge
vit-huge-patch14-clip-224.laion2b	Radford et al. (2021)	norm	Img-Txt	LAION2B	XLarge	ViT	TX	632.1M	xlarge
barlowtwins-rn50	Zbontar et al. (2021)	avgpool	SSL	IN1k	IN1k	ResNet	CNN	23.5M	small
dino-rn50	Caron et al. (2021)	avgpool	SSL	IN1k	IN1k	ResNet	CNN	23.5M	small
dino-vit-base-p16	Caron et al. (2021)	norm	SSL	IN1k	IN1k	ViT	TX	85.8M	small
dino-vit-small-p16	Caron et al. (2021)	norm	SSL	IN1k	IN1k	ViT	TX	21.7M	small
dino-xcit-medium-24-p16	Caron et al. (2021)	norm	SSL	IN1k	IN1k	ViT	TX	83.9M	small
dino-xcit-small-12-p16	Caron et al. (2021)	norm	SSL	IN1k	IN1k	ViT	TX	25.9M	small
dinov2-vit-base-p14	Oquab et al. (2024)	norm	SSL	LVD-142M	Large	ViT	TX	86.6M	small
dinov2-vit-giant-p14	Oquab et al. (2024)	norm	SSL	LVD-142M	Large	ViT	TX	1.1B	xlarge
dinov2-vit-large-p14	Oquab et al. (2024)	norm	SSL	LVD-142M	Large	ViT	TX	304.4M	large
dinov2-vit-small-p14	Oquab et al. (2024)	norm	SSL	LVD-142M	Large	ViT	TX	22.1M	small
jigsaw-rn50	Norooz & Favaro (2016)	avgpool	SSL	IN1k	IN1k	ResNet	CNN	23.5M	small
mae-vit-base-p16	He et al. (2022)	norm	SSL	IN1k	IN1k	ViT	TX	86.4M	small
mae-vit-huge-p14	He et al. (2022)	norm	SSL	IN1k	IN1k	ViT	TX	630.8M	xlarge
mae-vit-large-p16	He et al. (2022)	norm	SSL	IN1k	IN1k	ViT	TX	303.3M	large
mocov2-rn50	Chen et al. (2020c)	avgpool	SSL	IN1k	IN1k	ResNet	CNN	23.5M	small
pirl-rn50	Misra & van der Maaten (2020)	avgpool	SSL	IN1k	IN1k	ResNet	CNN	23.5M	small
rotnet-rn50	Gidaris et al. (2018)	avgpool	SSL	IN1k	IN1k	ResNet	CNN	23.5M	small
simclr-rn50	Chen et al. (2020a)	avgpool	SSL	IN1k	IN1k	ResNet	CNN	23.5M	small
swav-rn50	Caron et al. (2020)	avgpool	SSL	IN1k	IN1k	ResNet	CNN	23.5M	small
vicreg-rn50	Bardes et al. (2022)	avgpool	SSL	IN1k	IN1k	ResNet	CNN	23.5M	small
beit-base-patch16-224	Bao et al. (2022)	norm	Sup	IN21k + IN1k	IN21k	ViT	TX	86.5M	small
beit-base-patch16-224.in22k-ft-in22k	Bao et al. (2022)	norm	Sup	IN21k	IN21k	ViT	TX	102.6M	medium
beit-large-patch16-224	Bao et al. (2022)	norm	Sup	IN21k + IN1k	IN21k	ViT	TX	304.4M	large
beit-large-patch16-224.in22k-ft-in22k	Bao et al. (2022)	norm	Sup	IN21k	IN21k	ViT	TX	325.8M	large
convnext-base	Liu et al. (2022)	head.flatten	Sup	IN1k	IN1k	ConvNeXt	CNN	88.6M	small
convnext-large	Liu et al. (2022)	head.flatten	Sup	IN1k	IN1k	ConvNeXt	CNN	197.8M	medium
deit3-base-patch16-224	Touvron et al. (2021)	norm	Sup	IN1k	IN1k	ViT	TX	86.6M	small
deit3-base-patch16-224.fb-in22k-ft-in1k	Touvron et al. (2021)	norm	Sup	IN21k + IN1k	IN21k	ViT	TX	86.6M	small
deit3-large-patch16-224	Touvron et al. (2021)	norm	Sup	IN1k	IN1k	ViT	TX	304.4M	large
deit3-large-patch16-224.fb-in22k-ft-in1k	Touvron et al. (2021)	norm	Sup	IN21k + IN1k	IN21k	ViT	TX	304.4M	large
efficientnet-b3	Tan & Le (2019)	avgpool	Sup	IN1k	IN1k	EfficientNet	CNN	12.2M	small
efficientnet-b4	Tan & Le (2019)	avgpool	Sup	IN1k	IN1k	EfficientNet	CNN	19.3M	small
efficientnet-b5	Tan & Le (2019)	avgpool	Sup	IN1k	IN1k	EfficientNet	CNN	30.4M	small
efficientnet-b6	Tan & Le (2019)	avgpool	Sup	IN1k	IN1k	EfficientNet	CNN	43.0M	small
efficientnet-b7	Tan & Le (2019)	avgpool	Sup	IN1k	IN1k	EfficientNet	CNN	66.3M	small
resnet152	He et al. (2016)	avgpool	Sup	IN1k	IN1k	ResNet	CNN	60.2M	small
resnet50	He et al. (2016)	avgpool	Sup	IN1k	IN1k	ResNet	CNN	25.6M	small
resnext50-32x4d	He et al. (2016)	global_pool	Sup	IN1k	IN1k	ResNeXt	CNN	25.0M	small
seresnet50	Hu et al. (2018)	global_pool	Sup	IN1k	IN1k	SE-ResNet	CNN	28.1M	small
swin-base-patch4-window7-224	Liu et al. (2021)	global_pool	Sup	IN1k	IN1k	Swin-Transformer	TX	87.8M	small
swin-base-patch4-window7-224.ms-in22k	Liu et al. (2021)	global_pool	Sup	IN21k	IN21k	Swin-Transformer	TX	109.1M	medium
swin-large-patch4-window7-224	Liu et al. (2021)	global_pool	Sup	IN1k	IN1k	Swin-Transformer	TX	196.5M	medium
swin-large-patch4-window7-224.ms-in22k	Liu et al. (2021)	global_pool	Sup	IN21k	IN21k	Swin-Transformer	TX	228.6M	medium
vgg16	Simonyan & Zisserman (2015)	classifier.3	Sup	IN1k	IN1k	VGG	CNN	138.4M	medium
vgg19	Simonyan & Zisserman (2015)	classifier.3	Sup	IN1k	IN1k	VGG	CNN	143.7M	medium
vit-base-patch16-224	Dosovitskiy et al. (2021)	norm	Sup	IN1k	IN1k	ViT	TX	86.6M	small
vit-base-patch16-224.augreg-in21k	Dosovitskiy et al. (2021)	norm	Sup	IN21k	IN21k	ViT	TX	102.6M	medium
vit-huge-patch14-224.orig-in21k	Dosovitskiy et al. (2021)	norm	Sup	IN21k	IN21k	ViT	TX	630.8M	xlarge
vit-large-patch16-224	Dosovitskiy et al. (2021)	norm	Sup	IN1k	IN1k	ViT	TX	304.3M	large
vit-large-patch16-224.augreg-in21k	Dosovitskiy et al. (2021)	norm	Sup	IN21k	IN21k	ViT	TX	325.7M	large