BEYOND SCALARS: CONCEPT-BASED ALIGNMENT ANALYSIS IN VISION TRANSFORMERS

Johanna Vielhaben & Dilyara Bareeva & Jim Berend
Fraunhofer Heinrich Hertz Institute
Berlin, Germany
{johanna.vielhaben,dilyara.bareeva,jim.berend}@hhi.fraunhofer.de

Wojciech Samek[*]
Fraunhofer Heinrich Hertz Institute
BIFOLD – Berlin Institute for the Foundations of Learning and Data
Berlin, Germany
Technical University of Berlin
wojciech.samek@hhi.fraunhofer.de

Nils Strodthoff[†]
Carl von Ossietzky University of Oldenburg
Oldenburg, Germany
nils.strodthoff@uni-oldenburg.de

ABSTRACT

Vision transformers (ViTs) (Dosovitskiy et al., 2021) can be trained using various learning paradigms, from fully supervised to self-supervised. Diverse training protocols often result in significantly different feature spaces, which are usually compared through alignment analysis. However, current alignment measures quantify this relationship in terms of a single scalar value, obscuring the distinctions between common and unique features in pairs of representations that share the same scalar alignment. We address this limitation by combining alignment analysis with concept discovery, which enables a breakdown of alignment into single concepts encoded in feature space. This fine-grained comparison reveals both universal and unique concepts across different representations, as well as the internal structure of concepts within each of them. Our methodological contributions address two key prerequisites for concept-based alignment: 1) For a description of the representation in terms of concepts that faithfully capture the geometry of the feature space, we define concepts as the most general structure they can possibly form - arbitrary manifolds, allowing hidden features to be described by their proximity to these manifolds. 2) To measure distances between concept proximity scores of two representations, we use a generalized Rand index and partition it for alignment between pairs of concepts. We confirm the superiority of our novel concept definition for alignment analysis over existing linear baselines in a sanity check. The concept-based alignment analysis of representations from four different ViTs reveals that increased supervision correlates with a reduction in the semantic structure of learned representations.

1 INTRODUCTION

Vision Transformers are gaining increased popularity as backbones for various computer vision tasks. There is a large zoo of pre-trained models trained with various learning paradigms and a range of supervision strengths. To guide practitioners, previous work has evaluated performance on various common downstream tasks (Goldblum et al., 2023). A complimentary view of comparisons within and between models beyond quantitative accuracy is achieved by analyzing patterns in hidden

[*] corresponding author

[†] corresponding author

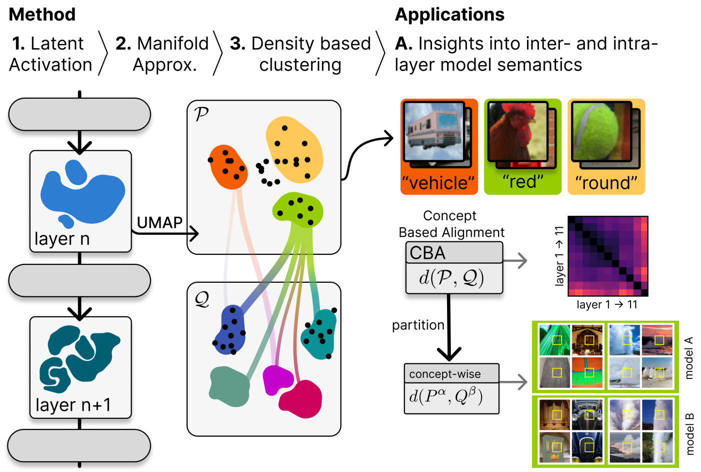

Figure 1: We combine concept discovery with alignment analysis for fine-grained insights into structures within and differences between latent activations. To this end, we investigate latent activations formed by intermediate layers, which according to the manifold hypothesis can be organized in terms of low-dimensional manifolds. We recover manifolds using density-based clustering applied to UMAP embeddings of the latent representations. The discovered structures in latent space do not only allow to characterize a single layer, but also the formation of structures between layers.

activations and measuring representational alignment between them (Raghu et al., 2021; Walmer et al., 2022).

When choosing a model for a downstream task, we want to understand how the model solves its pretraining task. Where does the model representation change the most and how? Which concepts, i.e. dominant structures in representation space, are encoded in lower layers vs. upper layers? Where does the model representation change the most and how? How structured are the representations? Does the model encode semantically similar concepts in spatial proximity to each other? How is the representation of model A different from that of model B across layers? Answering these questions can aid the selection of pre-trained models and the design of fine-tuning strategies through detailed insights into robustness and generalization capabilities. Previous work on alignment, however, only provides a single scalar value to measure alignment between representations at two different layers (Sucholutsky et al., 2023), leaving the questions above largely unanswered. In this paper, we propose a more fine-grained alignment analysis based on concepts that structure the latent representation. To this end, we represent the original activations by concept membership scores that quantify proximity to the discovered concepts. Then, we measure alignment between concept proximity scores of representations and can therefore partition it into the concepts. This gives insights into universal and specific concepts between representations of different layers or models, as well as how a single representation is structured.

To achieve concept-based alignment we need solutions for 1) concept discovery, and 2) measuring the alignment between concept proximity scores.

Previous work on concept discovery ranges from merely identifying neurons or other pre-existing units as concepts (Bau et al., 2017) to linear directions in feature space (Fel et al., 2023a). The most general definition so far relies on concepts as multi-dimensional linear subspaces (Vielhaben et al., 2023). The common strong assumption among these is the linearity of concept structures, which is challenging to verify and controversial (Bereska & Gavves, 2024; Csordás et al., 2024). For concepts that faithfully represent the underlying geometry of the representation, we avoid the linearity assumption and consider concepts as the most general structure they can form, namely as nonlinear manifolds. So far, alignment between representations has been measured as the similarity of similarities, e.g. through linear or kernel-based Centred Kernel Alignment (CKA) (Kornblith et al., 2019), which results in a single scalar value. Our fine-grained concept-based alignment measure requires a distance measure between concept proximity scores. Here, we choose a generalized Rand index between soft clusterings with pseudo metric properties (Hullermeier et al., 2012) that we partition into pairwise concept distances.

To summarize, our key idea is the following: We combine concept discovery with alignment analysis to provide insights into which concepts are universal or specific between two representations, and how structured a single representation is. We make the following methodological contributions to realize concept-based alignment: (1) We propose a novel concept definition of concepts as nonlinear manifolds to faithfully capture the geometry of the feature space with concept proximity scores. (2) We leverage a generalized Rand index with pseudo-metric properties to measure the alignment between concept proximity scores of two representations and partition it for fine-grained concept alignment.
We complement concept-based alignment analysis of ViTs trained under varying degrees of supervision from fully supervised to self-supervised with additional characteristics of concepts such as their intrinsic dimensionality. We find that representations of ViTs exhibit markedly different structures; specifically, increased supervision correlates with reduced semantic structure in the learned representations. This insight is crucial for understanding the model's reasoning processes and sheds light on the performance differences observed in quantitative analyses, such as those presented in the recent battle-of-the-backbones study (Goldblum et al., 2023).

2 Concept Discovery for Representational Alignment

This section is partitioned into three parts: First, we introduce our novel concept definition based on the manifold hypothesis. Then, we describe our methodology for discovering these concepts in latent activations, shown in Fig. 1. Finally, we describe how our concept-based description of hidden representations can be used to measure alignment between representations, identify commonalities and uniqueness between models, and investigate information flow within one model.

2.1 Concept definition

Motivation According to the manifold hypothesis, which is widely accepted in machine learning, many datasets, including image data that nominally lie in high dimensional space, can be described in terms of a few underlying latent factors and are thus concentrated on a (potentially disconnected) low-dimensional manifold embedded in high-dimensional space (Goodfellow et al., 2016). Naitzat et al. (2020) shows how a neural network trained on a toy classification problem solves the task by transforming the topology of the input data, and layerwise reducing the Betti numbers of the class-wise components. We hypothesize that state-of-the-art vision models behave similarly and try to recover the connected components in the hidden representations, which we call *concepts*.

Definition We analyze the hidden representation at an intermediate feature layer of a neural network. To this end, we split the model f into two parts, $f = g_l \circ h_l$, where h_l is the mapping to a hidden feature layer l. Our definition then relies on hidden representations $h_l(x_i) \in \mathbb{R}^{N' \times F}$ of input samples x_i from a set S. N' is the number of spatially separable elements in the representation, i.e. the number of tokens in a transformer model or the number of superpixels in a convolutional feature map. We spatially decompose the feature maps $h(x_i)$ into a set of $N = N' \cdot |S|$ feature vectors $\boldsymbol{\phi} \in \mathbb{R}^F$. Previously, concepts have been mostly defined as linear structures (Fel et al., 2023a; Zhang et al., 2021). The most general linear structure would be affine subspaces, which would already represent an extension compared to the recently considered definition as linear subspaces (Vielhaben et al., 2023). In this work, we generalize this idea even one step further and define concepts as manifolds in the F-dimensional feature space.

Definition 1 *We define a concept C^α, as a manifold in the d-dimensional feature space, represented by a point cloud $\{\boldsymbol{\phi}_j^\alpha\}$ consisting of the feature vectors $\boldsymbol{\phi}_j$ that lie on the concept manifold with index α.*

Benefits of concept manifold definition In the following, we want to compute concept proximity scores by which we measure alignment. Incorrect assumptions about the structure of the concept manifold, e.g., assuming it has no curvature (affine subspaces) or it is spherical and the distance to the manifold can be estimated by the distance to the centroid, directly lead to distorted concept proximity scores and hence to distorted alignment. Later, in a sanity check our definition performs best for measuring representational alignment.

2.2 Concept discovery

Clustering Having established our definition of concepts as manifolds in feature space, we now turn to the challenge of discovering these concepts through clustering. As stated above, we assume that feature vectors $\{\boldsymbol{\phi}_i\}$ from a hidden representation are sampled from a set of low-dimensional concept manifolds $\{C^\alpha\}$. Recovering these concept manifolds in high-dimensional space ($F = 768$ in our experiments) is a challenging clustering problem. Therefore, we revert to density-based clustering on a low-dimensional embedding of the data (Goh & Vidal, 2008; Herrmann et al., 2023). For this embedding, we utilize UMAP (Uniform Manifold Approximation and Projection) (McInnes & Healy, 2018), a dimensionality reduction technique that preserves local and some global structure. Given that we have no a priori knowledge about the number of clusters, we employ HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), which can handle clusters of varying densities (Campello et al., 2013). HDBSCAN builds a hierarchy of clusters based on density, represented by a condensed tree, and allows for robust handling of noise, making it suitable for the possibly intricate structure of feature representation spaces. While UMAP does not fully preserve density, its ability to maintain the overall structure of the data makes it a valuable preprocessing step before applying HDBSCAN. We use the HDBSCAN implementation from McInnes et al. (2017).

Concept proximity scores We leverage soft clustering with HDBSCAN based on the condensed tree which is roughly a density function over the data points to compute fuzzy cluster membership as described in (McInnes et al., 2017), which we formalize in the appendix for the reader's convenience. It is based on the distance to concept anchor points of each cluster and an outlier score, both derived from the condensed tree. We now have a fuzzy clustering $\mathcal{P}_{\{\boldsymbol{\phi}\}} = \{\boldsymbol{P}(\boldsymbol{\phi}_0), \ldots, \boldsymbol{P}(\boldsymbol{\phi}_N)\}$ with n clusters, where $\boldsymbol{P} \in [0,1]^n$ holds the concept proximity scores of each concept C^α. We interpret the concept proximity scores $P^\alpha(\phi)$ as the probability that a feature vector ϕ belongs to a concept P^α in clustering $\mathcal{P}$. This approach contrasts with previous concept assignment paradigms (Fel et al., 2023a; Vielhaben et al., 2023), which often rely on hard clustering, where each feature vector is assigned to a single concept, or linear methods that project onto specific concept directions, limiting the representation to a more rigid framework. In contrast, our soft clustering method allows for nuanced membership scores that reflect the degree of belonging to multiple concepts. In the following, we refer to our concept discovery method as *NLMCD* (non-linear multi-dimensional concept discovery).

2.3 Concept-based Representational Alignment

We now address the question of measuring representational alignment based on the concept proximity scores derived from fuzzy clustering.

Pseudo-metric between fuzzy clusterings The concepts are at this point characterized by a probabilistic clustering $\mathcal{P}_{\{\boldsymbol{\phi}\}} = \{\boldsymbol{P}(\boldsymbol{\phi}_0), \ldots, \boldsymbol{P}(\boldsymbol{\phi}_n)\}$, where $\boldsymbol{P}(\boldsymbol{\phi}_i) = [P^1(\boldsymbol{\phi}_i), \ldots, P^n(\boldsymbol{\phi}_i)]$. We want to measure the similarity between two probabilistic clusterings $\mathcal{P}, \mathcal{Q}$ from two different representations to evaluate how aligned their concepts are. For this purpose, we leverage an extension of the pair-based Rand index generalized to fuzzy clusterings proposed by Hullermeier et al. (2012). The original Rand index counts the number of concordant pairs (either two points are paired or not paired both clusterings) and disconcordant pairs (two points are paired in one clustering but not in the other). The distance between probabilistic clustering $\mathcal{P}, \mathcal{Q}$ is based on a generalized degree of concordance that is based on the *distance between two membership vectors* $d_{ms}(\boldsymbol{P}(\boldsymbol{\phi}_i), \boldsymbol{P}(\boldsymbol{\phi}_j))$:

$$d_{cross}(\mathcal{P}, \mathcal{Q}) = \frac{2}{n(n-1)} \sum_{i<j} |d_{ms}(\boldsymbol{P}(\boldsymbol{\phi}_i), \boldsymbol{P}(\boldsymbol{\phi}_j)) - d_{ms}(\boldsymbol{Q}(\boldsymbol{\phi}_i), \boldsymbol{Q}(\boldsymbol{\phi}_j))| \tag{1}$$

A commonly used choice for the distance d is $d_{ms}(\boldsymbol{P}(\boldsymbol{\phi}_i), \boldsymbol{P}(\boldsymbol{\phi}_j)) = 1 - ||\boldsymbol{P}(\boldsymbol{\phi}_i) - P(\boldsymbol{\phi}_j)||_1$ (DeWolfe & Andrews, 2023). Finally, we refer to the similarity between two clusterings, derived from the uncovered concepts, as *Concept-Based Alignment* (CBA):

$$\text{CBA} = 1 - d_{cross}(\mathcal{P}, \mathcal{Q}) \tag{2}$$

We choose this measure because $d_{cross}(\mathcal{P}, \mathcal{Q})$ is a pseudo-metric satisfying desirable properties[1] that ease interpretation Also, when $\mathcal{P}, \mathcal{Q}$ are crisp partitions, CBA reduces to the original Rand index.

[1] 1) Identity: $d(x,x) = 0$ for all x, 2) Symmetry: $d(x,y) = d(y,x)$ for all x, y, 3) Triangle Inequality: $d(x,z) \leq d(x,y) + d(y,z)$ for all x, y, z.

Similarity index between clusters In contrast to conventional measures for representational alignment that yield a single scalar value, our approach provides a more nuanced measure of representational alignment by assessing similarities and differences between clusters. To measure distance between two clusters P^α, Q^β from two clusterings $\mathcal{P}$, $\mathcal{Q}$, we decompose the distance in Eq. 1 into the contribution of single concepts P^α, Q^β and measure the *pairwise similarity between the membership scores* of each feature

$$d_{cross}(P^\alpha,Q^\beta) = \tfrac{2}{n(n-1)} \sum_{i<j} ||P^\alpha(\boldsymbol{\phi}_i) - P^\alpha(\boldsymbol{\phi}_j)| - |Q^\beta(\boldsymbol{\phi}_i) - Q^\beta(\boldsymbol{\phi}_j)|| \tag{3}$$

Due to the absolute value in Equation (1), summing over all pairs α, β does not yield the total $d_{cross}(\mathcal{P}, \mathcal{Q})$, but by the triangle inequality $\sum_{\alpha,\beta} d_{cross}(P^\alpha, Q^\beta) \geq d_{cross}(\mathcal{P}, \mathcal{Q})$ the sum is an upper bound for the overall distance between two clusterings.

3 Related work

Alignment Representational alignment measures are categorized, with a particular emphasis on Centered Kernel Alignment (CKA) in (Kornblith et al., 2019). CKA evaluates the similarity of similarities, either linearly or under a non-linear kernel. Similarly, (Doimo et al., 2020) measure alignment through the similarities of binary k-nearest neighbor adjacency matrices, which resembles CKA with a narrow Gaussian kernel. Our method relates to CKA in that it condenses these similarities into clusters and subsequently measures the similarity between these clusterings.

Concept discovery Most existing methods model concepts as linear directions (Ghorbani et al., 2019; Zhang et al., 2021; Fel et al., 2023b;a). Generalizing this definition, (Vielhaben et al., 2023) suggest that concepts can be represented more faithfully as multidimensional linear subspaces, which they discover through sparse subspace clustering. While above methods operate unsupervised without concept labels, Crabbé & van der Schaar (2022) employ kernel classifier for supervised, nonlinear concept discovery, showing improvement over linear concepts. In the field of *mechanistic interpretability*, many studies aim to enumerate all *features* encoded in the representations of neural networks (Bricken et al., 2023). This line of work focuses mainly on language models, often identifying one-dimensional linear features using sparse autoencoders (SAEs) (Marks et al., 2024; Huben et al., 2024; Gao et al., 2024). Lan et al. (2025) find high representational alignment between SAE representations of different language models.However, Csordás et al. (2024) find evidence for the existence of non-linear features. Unlike these approaches, our main goal in concept discovery is representation summarization for alignment measurement, rather than interpretability or feature enumeration. For this reason, we employ the most general, non-linear concept definition.

Comparison of Vision Models On the one hand, alignment measures such as CKA have been used to compare the representations of various architectures, including ViTs and ResNets trained on different tasks, together with the analysis of patterns in attention maps (Walmer et al., 2022; Raghu et al., 2021). Further, the analysis of attention patterns reveals differences between self-supervised ViTs (Park et al., 2023). On the other hand, downstream performance is analyzed to guide the selection of pre-trained models for transfer learning. Through this, (Kornblith et al., 2018) shows that models pre-trainined on ImageNet generalize well but when used as feature extractors in transfer learning, i.e. when weights are completely frozen, perform badly in some settings, suggesting that the features of the last layers do not generalize well. An extensive evaluation of the downstream performance of a large selection of vision models on classification, detection, image retrieval, and generalization is available in (Goldblum et al., 2023).

4 Results

We evaluate concept discovery in Sec. 4.1, check the superiority of our new concept definition over linear baselines for concept alignment analysis in Sec. 4.2, and perform a concept-alignment analysis between four ViTs in Sec. 4.3.

4.1 Concept discovery

First, we outline the concept discovery procedure as described in Sec. 2.2 and evaluate the quality of the UMAP embeddings used for HDBSCAN clustering and the clustering itself. For concept

Table 1: Pre-trained models we study with concept-based alignment, which range from fully supervised to text-image contrastive to self-supervised. Sources for the model weights are in the appendix.

Model	Training Data	Training Task
FS (Steiner et al., 2022)	ImageNet-1k (Russakovsky et al., 2015)	Fully supervised learning with labeled data for classification task.
CLIP (Radford et al., 2021)	WebImageText (Radford et al., 2021)	Contrastive learning between images and text.
DINO (Caron et al., 2021)	ImageNet-1k	Knowledge distillation enforcing consistency between augmented views of the same image.
MAE (He et al., 2022)	ImageNet-1k	Masked autoencoders to reconstruct missing pixels of input data.

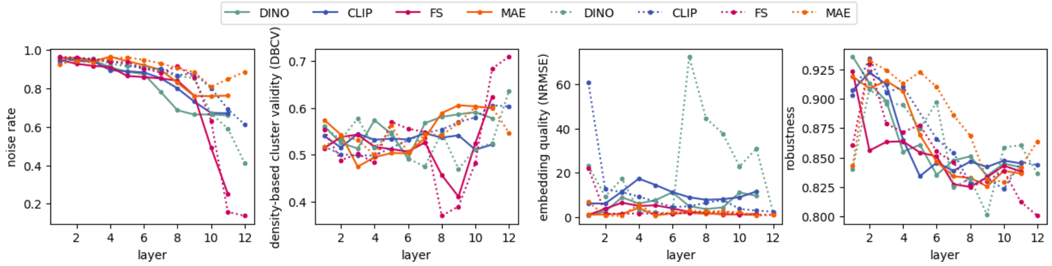

Figure 2: Quality of concept discovery: NRMSE measures the RMSE between the distance matrix of the original and embedded activations, normalized by the average distance in the original embedding, and shows how faithfully the UMAP embedding captures the geometry of the representation. DBCV is a density-based clustering validity index that contrasts intra- vs inter-cluster density with scores in $[-1, 1]$ where higher is better. The noise rate is the ratio of points classified as noise in HDBSCAN. Robustness is measured between two runs by concept-alignment from Eq. 2. Results are across layers for CLS (dotted) and SEQ (solid) token representations of the models in Tab. 1.

discovery and later analysis of representational alignment, we use a random subset of 25 % of the ImageNet train set, stratified samples across all 1000 classes. We study four different ViTs (Dosovitskiy et al., 2021) with the same architecture (base, patch size 16, input size 224) but different training objectives and training datasets described in Tab. 1. We perform concept discovery separately for the sequence (SEQ) and the CLS token. We extract activations at the last MLP layer of each of the twelve transformer blocks. Due to computational constraints of UMAP and HDBSCAN limiting the number of tokens for clustering, we reduce the 196 SEQ tokens per image to a single representative token, which also facilitates comparison with CLS experiments. We obtain this token by average pooling over a central 4×4 block of tokens, assuming the image center contains more diverse concepts, while peripheral regions may predominantly capture repetitive background elements. Further, for SEQ tokens, we discard the last block as for the considered models only the CLS token in the final layer enters the loss. To evaluate concept discovery, first, we report the rate of points classified as noise by HDBSCAN. Second, to evaluate how well the HDBSCAN clusters are separated, we compute a density-based validity index (DBCV) Moulavi et al. (2014), which measures intra- vs inter-cluster density and yields scores in $[-1, 1]$ where higher is better.[2] Third, we evaluate how well the embedding on which we perform the clustering preserves the distances by measuring the root mean squared error between the distance matrices in the original representation and its UMAP embedding, normalized by the mean pair-wise distances in the original embedding (NRMSE). Lastly, we evaluate how robust our approach is by measuring the alignment between two runs with different initializations by CBA from Eq. 2. We detail the hyperparameter tuning for UMAP and HDBSCAN, which is based on DBCV, and the final hyperparameters used in all sub-

[2] To treat the noise rate and validity index separately, we do not weight the average for the DBCV across clusters by the cluster size as proposed in Moulavi et al. (2014).

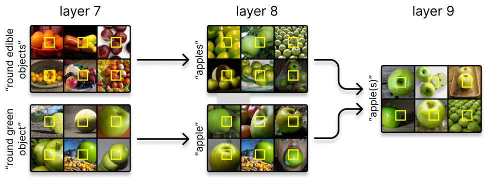

Figure 3: Concept formation graph for the concept “apple(s)” in layer 9 of the FS model. Each concept is represented by six randomly sampled images containing a token assigned to that concept (highlighted in a yellow frame).

sequent experiments, in the appendix. Turning to the results on embedding and clustering quality presented in Fig. 2, we observe that NRMSE is consistently low for most layers and models, only for CLIP CLS representations in layer one and DINO CLS representations between layer six and eleven it deviates to high values. The density-based validity ranges roughly between 0.4 and 0.7 indicating medium clustering quality, but is similar across models and SEQ tokens vs. CLS tokens. Given how challenging the clustering task is, we view this result as decent and refer to the convincing qualitative impression of the clusters in Fig. 3 and Fig. 4. Noise rates are rather high but decrease across layers. The high noise may be due to insufficiently dense sampling, i.e. thorough sampling of noisy regions could result in concept clusters. Robustness decreases for all models across layers but stagnates at around 0.84 for most models in the late layers. For the qualitative evaluation of our concept discovery method, we construct *concept formation graphs* (CFGs) that depict the flow of token assignments to concepts from one layer to the next as an unweighted, directed graph. Fig. 3 displays the formation of the “apples” concept throughout the layers of the FS model. Note that these graphs may be incomplete, as some nodes might not be detected by the clustering method, illustrating the under-sampling problem described above. Additional examples for other models and the detailed algorithm for CFG construction are provided in the appendix.

4.2 Sanity checking concept structure for alignment and CKA comparison

We use a sanity check to demonstrate how concept-based alignment analysis benefits from concepts defined as non-linear manifolds by comparing against concept alignment based on other definitions and discovery methods. Additionally, we compare CBA from Eq. 2 against CKA as an established representational alignment measure that offers a single scalar to indicate alignment. The sanity check is based on the assumption that adjacent layers should be more aligned than distant ones. For each layer, we compute the weighted Kendall’s Tau correlation (Vigna, 2015) between alignment scores and layer distances separately for upstream and downstream layers. We apply hyperbolic weights to prioritize ranking closer layers correctly, as their alignment is more meaningful than that of distant layers. Separating upstream and downstream layers accounts for the possibility that representations can change at different rates (e.g., layer six may be less aligned with layer seven than layer four) while also avoiding tied ranks. Averaging all these correlations across model layers provides a check on whether the alignment measure reflects expected structural relationships. We compare NLMCD concepts against one-dimensional linear subspaces discovered by PCA (Zhang et al., 2021; Fel et al., 2023a), multi-dimensional linear subspaces discovered by MCD (Vielhaben et al., 2023), and spherical non-linear concepts discovered by KMeans clustering (Fel et al., 2023a). To obtain soft concept membership scores for the linear subspaces, we project the feature vector onto the concept subspace and clip to negative values to 0, as we argue that a feature vector pointing into the opposite direction of a concept signifies the concept not being active. For KMeans concepts, we measure concept proximity by the euclidean distance to the cluster centroid. We also normalize concept membership scores $P^{\alpha'} = P^{\alpha} / \sum_{\alpha} P^{\alpha}$ as their sum is required to be less bounded by one $\sum_{\alpha} P^{\alpha} \leq 1$ in Eq. 1. There is no direct way to estimate the number of concepts for PCA, MCD and KMeans, so we use all $F = 768$ components for PCA for a conservative baseline, and the number of concepts discovered by NLMCD for MCD and KMeans discovery. We present the sanity check in Tab. 2 for SEQ and CLS token alignment. We find that for SEQ tokens, NLMCD shows higher scores than linear concepts except for DINO, where PCA can match its performance. Further, only for DINO and MAE simple nonlinear KMeans is en par with NLMCD. Similarly, for CLS tokens, NLMCD mostly outperforms other concept methods, while only for FS can PCA, and for CLIP and MAE can KMeans match its performance. Further, regarding the comparison to CKA, for SEQ

Table 2: Sanity check for concept alignment, based on weighted Kendall Tau (Vigna, 2015) between alignment and layer distance. We compare the suitability of NLMCD for CBA concepts against other methods: one-dimensional linear subspaces (PCA), multi-dimensional linear subspaces (MCD), and spherical non-linear concepts (KMeans). Additionally, we compare CBA against CKA. Results within the same standard error interval as the top score for each model are **bold** and those CBA results within the same interval as NLMCD-CBA are *italic*. NLMCD consistently outperforms other concept approaches. While NLMCD-CBA and CKA are en-par, CBA offers the advantage of fine-grained concept-based alignment.

		FS	CLIP	DINO	MAE
SEQ	PCA-CBA	0.91	0.91	*0.88*	0.84
	MCD-CBA	0.90	0.92	0.85	0.87
	KMeans-CBA	0.94	0.82	*0.87*	***0.98***
	NLMCD-CBA	***0.97***	***0.98***	*0.92*	***0.98***
	CKA	**0.98**	0.94	**0.99**	**0.99**
CLS	PCA-CBA	***0.92***	0.91	0.78	0.78
	MCD-CBA	0.82	0.73	0.62	0.73
	KMeans-CBA	0.86	***0.96***	0.73	***0.89***
	NLMCD-CBA	***0.93***	***0.96***	***0.91***	***0.94***
	CKA	**0.93**	**0.97**	**0.89**	**0.93**

and CLS, NLMCD-CBA performs en par with CKA, while offering the significant advantage of providing fine-grained concept-based alignment.

4.3 Concept Alignment Analysis

We now investigate concept-based alignment described in Sec. 2.3 between representations across layers and models. We structure the analysis into *intra-model* and *inter-model*. Due to limited space, we focus on SEQ representation and defer the CLS representation analysis to the appendix.

Intra-model representations We analyze how representations are transformed within one model and how they are structured across layers. To supplement concept-based alignment analysis between representations, we further evaluate alignment of concepts with labels from ImageNet-1k, concept count, and the intrinsic dimensionality of each concept. With this analysis, we address the questions raised in the introduction: 1) Where does the model representation change the most and how? 2) Which concepts are encoded in lower layers vs. upper layers? 3) How structured are the latent representations - does the model encode semantically similar concepts in spatial proximity to each other?

Where does the model representation change the most and how? First, we focus on the intra-model alignment heatmaps between SEQ representations across layers measured by CBA from Eq. 2 in the upper row of Fig. 4. Interestingly, the transformation process in CLIP, DINO and MAE models is split between the first, i.e., layer one to six, and the second model half, i.e., layer six to eleven. The concept characteristics in Fig. 5 reflect this break and give insight into how the representation is transformed between the break from layers six and seven. The concept count increases rather smoothly across layers for these models, but picks up at layer seven. In contrast, class label alignment has a marked increase at this point. For DINO and MAE, the average intrinsic dimensionality of concepts slightly decreases at this point but increases further for CLIP. In contrast to the models above, the FS model exhibits a pronounced change rate between nine and ten, resulting in a sudden enhancement in class alignment at layer 10, accompanied by a marked increase in the number of clusters and intrinsic dimensionality. This is reflected in a low alignment between representations in the last two blocks of the FS model and indicates a nucleation process, where concepts begin to separate into distinct classes used for supervised training. This nucleation process has been previously observed in ResNets (Doimo et al., 2020).

Which concepts are encoded in lower layers vs. upper layers? How structured are the representations? We now zoom in and partition the representation into single concepts, at layer six just before the block separation in CLIP, DINO, MAE and at layer eleven as the last layer of the second model part. We construct a UMAP embedding based on the distance between concept pairs measured by $d_{cross}(P^{\alpha}, P^{\beta})$ from Eq. 3. Each point in this *concept atlas* corresponds to a different

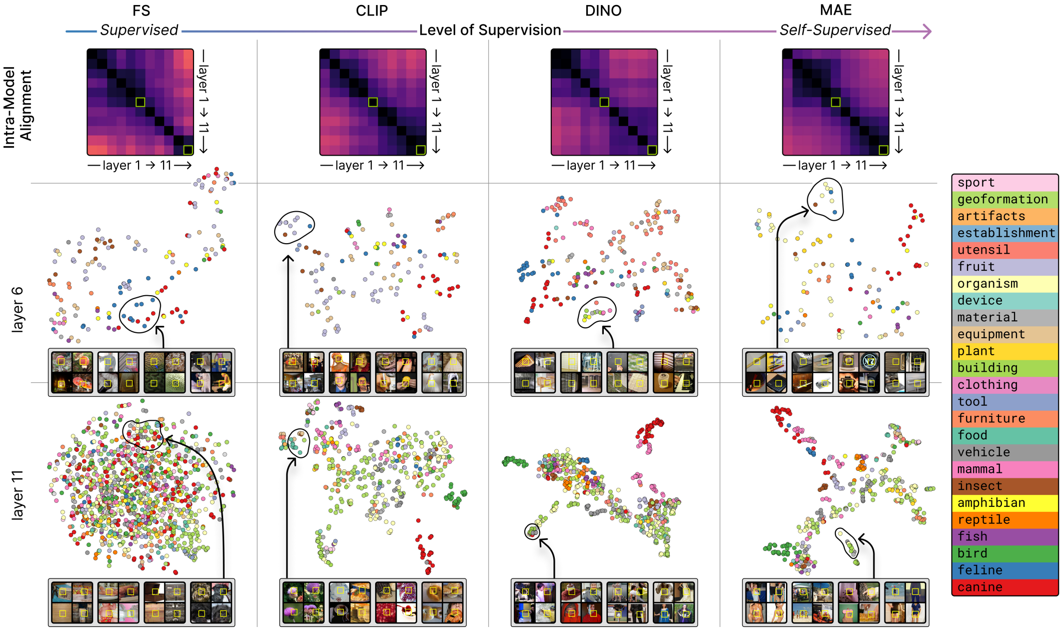

Figure 4: Intra-model relationships of SEQ representations. **Upper row:** We show CBA from Eq. 2 to visualize how representations are transformed across layers of the models from Tab. 1 (darker pixels correspond to higher alignment). We observe a nucleation process between layer 9 and 10 in FS and smoother processing split into two major blocks between layer 1-6 and 6-11 in CLIP, DINO and FS. **Center and bottom row:** we zoom into the representations at layer 6 and 11 and partition the scalar CBA alignment into single concepts. We show a UMAP embedding constructed from the pairwise distance of concept measured by $d_{cross}(P^{\alpha}, P^{\beta})$ from Eq. 3. Each point in this *concept atlas* corresponds to a distinct concept P^{α}. To convey their meaning, for some concepts, we show four random input tokens from the members of the concept cluster P^{α} marked by a yellow box in the entire image. The stronger the supervision during ViT training ranging from FS, over CLIP to DINO and MAE, the less semantically organized are the representations at layer 11.

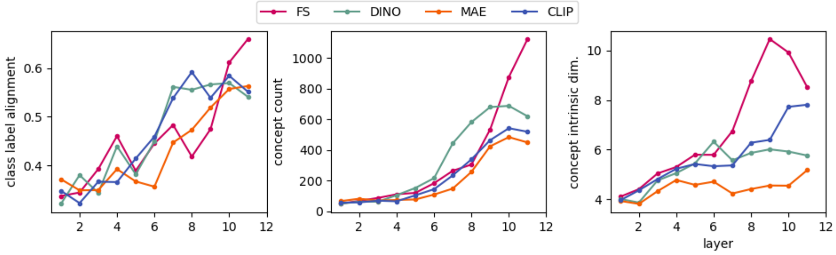

Figure 5: To supplement the intra-model alignment analysis, we evaluate alignment between concepts and ImageNet-1k class labels (based on CBA from Eq. 2), concept count, and the average intrinsic dimensionality (based on (Facco et al., 2017)) across concepts.

concept P^{α}. To convey their meaning, we show four random input tokens from the members of the concept cluster P^{α} (framed by a yellow box). Concept atlases for the representation at layer six and eleven across all models in give a visual impression of how semantically organized the concepts are. To guide the eye, we color-code the concept clusters based on categories derived from the ImageNet-1k labels of the images from which the patches were extracted. We first map these labels to more abstract categories[3] using the WordNet hierarchy (Miller, 1995). After mapping, we perform a majority vote among all patches in a cluster to assign the category. At layer six, concepts

[3] The mapping is provided in the appendix. Note that this labeling is only a proxy and may not accurately reflect the actual content of the patches—for instance, a patch might show grass on which an animal stands.

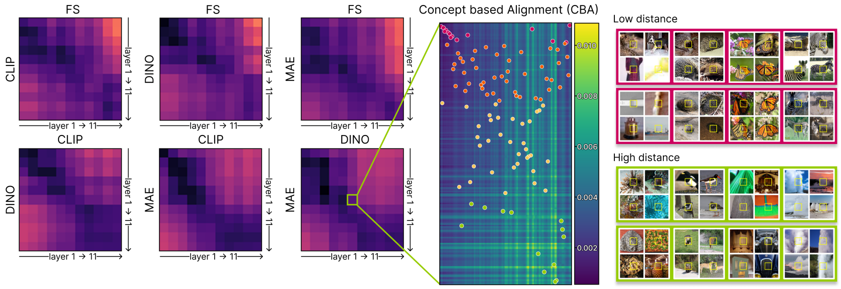

Figure 6: Inter-model relationships of SEQ representations. **Left:** We show CBA from Eq. 2 to visualize how representations differ between the models in Tab. 1 (darker pixels correspond to higher alignment). **Right:** We zoom into the concept-wise distances $d_{cross}(P^{\alpha}, P^{\beta})$ from Eq. 3 between the representation of layer six in MAE and DINO. To select specific examples of concpet pairs for inspection, we match concepts via the Hungarian algorithm that minimizes the sum of distances of pairs and mark these pairs with points colored to indicate concept distance (magenta for low and green for high distance). We illustrate some example pairs with low and high distance.

appear structured, but not yet aligned with the WordNet categories. By visual inspection, concepts are less abstract and rather encode structures, shapes and object parts. Representations across ViTs at this layer show a similar level of structuredness. In contrast, at layer eleven, the FS representation is notably less semantically organized than that of the other models. For CLIP, DINO, and MAE we point out how well the canine concepts are separated. To further exemplify, human body parts like neck, shoulder, and legs are grouped together in the representation of DINO and MAE. This alignment requires not only the preservation of local, or intra-cluster distances, but also the maintenance of broader, inter-cluster distances. We conclude that supervised training for the FS model does not enforce this level of semantical organization. In fact, it might make sense to push similar concepts apart in feature space to avoid confusion. However, this likely has negative implications for generalization to other tasks.

Inter-model relations We now analyze how the representations between two different models differ and present CBA from Eq. 2 between all layers of the models from Tab. 1 in the upper part of Fig. 6. We observe higher alignment between the self-supervised models DINO and MAE than with CLIP and the FS model in the alignment heatmaps. Further, layers of the first are more aligned than those of the second half across all models pairs. We conclude that basic foundational features are learned similarly across models, while later layers diverge as the models specialize to concepts serving their pre-training task.

How is the representation of model A different from that of model B? We zoom in into the distance $d_{cross}(P^{\alpha}, P^{\beta})$ from Eq. 3 between concept pairs from representations of DINO and MAE layer six in the lower part of Fig. 6. We visualize the distance matrix between all concepts. To select specific examples of pairs for inspection, we match concepts via the Hungarian algorithm that minimizes the sum of distances of pairs. From this selection, we show concept pairs with low distance (blurriness, satchel of a hedgehog, zebra stripes) and high distance (complex high-frequency structure, vertically textured structure, fountain and fog-like). Both the universal concepts with low distance and the more specific concepts with high distance seem to correspond mainly to structure and texture but visual discrepancy is more pronounced for high-distance concepts.

5 Conclusion

We propose a novel approach that combines concept discovery with representational alignment analysis in ViTs. With concept-based alignment analysis, we answer the questions raised in the introduction and examine the structuredness in feature spaces of different ViTS, as well as fine details between the concepts of two different models. These insights are not available through traditional scalar alignment measures. Understanding the structured nature of latent spaces can guide practitioners in choosing models that not only perform well on benchmark datasets but also exhibit robust feature representations for downstream tasks. For instance, the nucleation process in FS emphasizes the importance of model structure over mere classification accuracy when selecting a pre-trained model.

Limitations The computational scalability of HDBSCAN limits the sampling of feature vectors which makes undersampled concept regions appear as noise. The limited variability of ImageNet-1k might obfuscate the meaning of a concept, e.g. when a concept represents a color but there are only dog patches of that color.

Acknowledgments This work was supported by the European Union's Horizon Europe research and innovation programme (EU Horizon Europe) as grants [ACHILLES (101189689), TEMA (101093003)]; and the German Research Foundation (DFG) as research unit DeSBi [KI-FOR 5363] (project-id: 459422098).

REFERENCES

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6. Survey Certification, Expert Certification.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2013. URL https://api.semanticscholar.org/CorpusID:32384865.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Jonathan Crabbé and Mihaela van der Schaar. Concept activation regions: A generalized framework for concept-based explanations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 2590–2607. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/11a7f429d75f9f8c6e9c630aeb6524b5-Paper-Conference.pdf.

Róbert Csordás, Christopher Potts, Christopher D Manning, and Atticus Geiger. Recurrent neural networks learn to store and generate sequences using non-linear representations. In *The 7th BlackboxNLP Workshop*, 2024. URL https://openreview.net/forum?id=NUQeYgg8x4.

Ryan DeWolfe and Jeffery L. Andrews. Random models for fuzzy clustering similarity measures. *ArXiv*, abs/2312.10270, 2023. URL https://api.semanticscholar.org/CorpusID:266348826.

Diego Doimo, Aldo Glielmo, Alessio Ansuini, and Alessandro Laio. Hierarchical nucleation in deep neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pp. 7526–7536, Red Hook, NY, USA, December 2020. Curran Associates Inc. ISBN 9781713829546.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, Sep 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-11873-y. URL https://doi.org/10.1038/s41598-017-11873-y.

Thomas Fel, Victor Boutin, Louis Béthune, Remi Cadene, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 54805–54818. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/abf3682c9cf9245a0294a4bebe4544ff-Paper-Conference.pdf.

Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2711–2721, June 2023b.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf.

Alvina Goh and Rene Vidal. Clustering and dimensionality reduction on Riemannian manifolds. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, June 2008. doi: 10.1109/CVPR.2008.4587422. ISSN: 1063-6919.

Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, Ramalingam Chellappa, Andrew Gordon Wilson, and Tom Goldstein. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. *ArXiv*, abs/2310.19909, 2023. URL https://api.semanticscholar.org/CorpusID:264818042.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Moritz Herrmann, Daniyal Kazempour, Fabian Scheipl, and Peer Kröger. Enhancing cluster analysis via topological manifold learning. *Data Mining and Knowledge Discovery*, September 2023. ISSN 1573-756X. doi: 10.1007/s10618-023-00980-2. URL https://doi.org/10.1007/s10618-023-00980-2.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Eyke Hullermeier, Maria Rifqi, Sascha Henzgen, and Robin Senge. Comparing fuzzy partitions: A generalization of the rand index and related measures. *IEEE Transactions on Fuzzy Systems*, 20 (3):546–556, 2012. doi: 10.1109/TFUZZ.2011.2179303.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, 2018. URL https://api.semanticscholar.org/CorpusID:43928547.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Sparse autoencoders reveal universal feature spaces across large language models, 2025. URL https://arxiv.org/abs/2410.06981.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models, 2024. URL https://arxiv.org/abs/2403.19647.

Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*, abs/1802.03426, 2018. URL https://api.semanticscholar.org/CorpusID:3641284.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), mar 2017. doi: 10.21105/joss.00205. URL https://doi.org/10.21105%2Fjoss.00205.

George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219748. URL https://doi.org/10.1145/219717.219748.

Davoud Moulavi, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. Density-based clustering validation. In *SDM*, 2014. URL https://api.semanticscholar.org/CorpusID:16656312.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *The Journal of Machine Learning Research*, 21(1):184:7503–184:7542, January 2020. ISSN 1532-4435.

Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? *ArXiv*, abs/2305.00729, 2023. URL https://api.semanticscholar.org/CorpusID:258426737.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *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.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:237213700.

Sebastian Raschka, Joshua Patterson, and Corey Nolet. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. 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.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew Kyle Lampinen, Klaus-Robert Muller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *ArXiv*, abs/2310.13018, 2023. URL https://api.semanticscholar.org/CorpusID:264405712.

Johanna Vielhaben, Stefan Bluecher, and Nils Strodthoff. Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees. *Transactions on Machine Learning Research*, January 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=KxBQPz7HKh&noteId=QlqzbYGWI3.

Sebastiano Vigna. A weighted correlation index for rankings with ties. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pp. 1166–1176, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741088. URL https://doi.org/10.1145/2736277.2741088.

Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7486–7496, 2022. URL https://api.semanticscholar.org/CorpusID:254366577.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11682–11690, May 2021. doi: 10.1609/aaai.v35i13.17389. URL https://ojs.aaai.org/index.php/AAAI/article/view/17389.

A Appendix

B HDBSCAN

After concept discovery with HDBSCAN, we compute concept proximity scores $\mathcal{P}_{\{\boldsymbol{\phi}\}} = \{\boldsymbol{P}(\boldsymbol{\phi}_0), \ldots, \boldsymbol{P}(\boldsymbol{\phi}_N)\}$,$\boldsymbol{P} \in [0,1]^n$ holds the concept membership scores $P^\alpha(\boldsymbol{\phi})$ of each concept C^α. These rely on the implementation of soft clustering with HDBSCAN from (McInnes et al., 2017), which we formalize here for the reader's convenience.

Clustering HDBSCAN first transforms the feature space using a density-informed metric called *mutual reachability distance*

$$\begin{aligned} \text{MRD}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) = \max(&\text{coreDistance}_k(\boldsymbol{\phi}_i), \\ &\text{coreDistance}_k(\boldsymbol{\phi}_j), d(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j)) \end{aligned} \tag{4}$$

where coreDistance$_k(\boldsymbol{\phi})$ is the distance between a point $\boldsymbol{\phi}$ and its k-nearest neighbor. Based on the mutual reachability distance between all pairs, a minimum spanning tree is constructed that connects all points and minimizes the sum of the edges weighted by MRD. From this, a hierarchical tree is constructed via robust single linkage clustering. The hierarchical tree is condensed by eliminating insignificant clusters and simplifying the hierarchy. This is achieved by selecting a range of *persistence* values λ, which are the inverses of the mutual reachability distances ($\lambda = 1/\text{MRD}$). Clusters that persist over significant ranges of λ, i.e. they are stable across multiple density levels, are retained, while clusters that exist only over narrow ranges of λ are considered noise and pruned from the tree. The result is a condensed tree that focuses on the most significant clusters. Finally clusters are extracted from the condensed tree either based on their stability across different density levels or simply the leaf nodes are identified as clusters.

Soft clustering with HDBSCAN The soft cluster membership scores combine a distance-based membership with and an outlier score.

For the distance-based membership to cluster C^α, first k exemplar points $\{\boldsymbol{\phi}_i^\alpha\}$, $i \in [1, k]$, are extracted. A single centroid is not enough to characterize a cluster as its shape can be arbitrary. The exemplar points are the points within the leaf nodes beneath cluster C^α with maximum persistence λ in the condensed tree, i.e. the densest points where the cluster persists. Then, the distance membership score between a point $\boldsymbol{\phi}$ and a cluster C^α is the inverse minimum distance across the exemplar points $\{\boldsymbol{\phi}_i^\alpha\}$,

$$M^\alpha(\boldsymbol{\phi})_{\text{dist}} = \frac{[\min_i(d(\boldsymbol{\phi}, \boldsymbol{\phi}_i^\alpha))]^{-1}}{\sum_\beta [\min_j(d(\boldsymbol{\phi}, \boldsymbol{\phi}_j^\beta))]^{-1}}\,, \tag{5}$$

normalized across all clusters. The outlier-based membership compares a point's membership persistence to the total persistence of a cluster:

$$M^\alpha(\boldsymbol{\phi})_{\text{membership}} = \frac{\lambda_{\boldsymbol{\phi}\to C^\alpha} - \lambda_{\text{birth}}^{C^\alpha}}{\lambda_{\text{max}}^{C^\alpha} - \lambda_{\text{birth}}^{C^\alpha}}\,. \tag{6}$$

Here, $\lambda_{\text{birth}}^{C^\alpha}$ is the persistence value at which cluster C^α first appears, i.e. its birth point in the condensed tree and $\lambda_{\boldsymbol{\phi}\to C^\alpha}$ is the persistence value at which point $\boldsymbol{\phi}$ would join cluster C^α. Finally, distance and outlier-based membership are combined with stronger emphasis on outlier-based membership,

$$M^\alpha(\boldsymbol{\phi}) = (M^\alpha(\boldsymbol{\phi})_{\text{dist}})^{1/2} \cdot (M^\alpha(\boldsymbol{\phi})_{\text{membership}})^2\,, \tag{7}$$

and normalized $M_{\text{norm}}^\alpha(\boldsymbol{\phi}) = M^\alpha(\boldsymbol{\phi}) / \sum_\beta M^\beta(\boldsymbol{\phi})$. This membership score $M_{\text{norm}}^\alpha(\boldsymbol{\phi})$ can be interpreted as the probability that a point $\boldsymbol{\phi}$ belongs to cluster C^α, given that the point belongs to some cluster,

$$M_{\text{norm}}^\alpha(\boldsymbol{\phi}) \equiv P(\boldsymbol{\phi} \in C^\alpha \mid \exists\beta : \boldsymbol{\phi} \in C^\beta)\,. \tag{8}$$

We want to compute the joint probability $P(\boldsymbol{\phi} \in C^\alpha)$, which includes the probability that $\boldsymbol{\phi}$ may be noise,

$$P(\boldsymbol{\phi} \in C^\alpha) = P(\boldsymbol{\phi} \in C^\alpha \mid \exists\beta : \boldsymbol{\phi} \in C^\beta) P(\exists\beta : \boldsymbol{\phi} \in C^\beta)\,. \tag{9}$$

Here, $P(\exists\beta : \boldsymbol{\phi} \in C^\beta)$ is the probability that $\boldsymbol{\phi}$ belongs to some cluster. To estimate $P(\exists\beta : \boldsymbol{\phi} \in C^\beta)$, the λ value at which $\boldsymbol{\phi}$ would join the nearest cluster is compared to the maximum λ value of that cluster,

$$P(\exists\beta : \boldsymbol{\phi} \in C^\beta) = \frac{\lambda_{\boldsymbol{\phi}\to C^\alpha}}{\lambda_{\text{max}}^{C^\alpha}}\,, \tag{10}$$

where $\lambda_{\boldsymbol{\phi}\to C^\alpha}$ is the persistence value at which point $\boldsymbol{\phi}$ would join its nearest cluster C^α and $\lambda_{\text{max}}^{C^\alpha}$ is the maximum λ value of cluster C^α. Thus, the final probability, that point $\boldsymbol{\phi}$ belongs to cluster C^α is,

$$P^\alpha(\boldsymbol{\phi}) = \frac{\lambda_{\boldsymbol{\phi}\to C^\alpha}}{\lambda_{\text{max}}^{C^\alpha}} \cdot M_{\text{norm}}^\alpha(\boldsymbol{\phi})\,. \tag{11}$$

C Details on experimental setup

Here, we provide further details on the experiments.

ViT sources We list the URL of each Vision Transformer provided by the timm library (Wightman, 2019):

- FS: `https://huggingface.co/timm/vit_base_patch16_224.augreg_in1k`
- CLIP: `https://huggingface.co/timm/vit_base_patch16_clip_224.openai`
- DINO: `https://huggingface.co/timm/vit_base_patch16_224.dino`
- MAE: `https://huggingface.co/timm/vit_base_patch16_224.mae`

Hyperparameters for UMAP and HDBSCAN We tune hyperparameters of UMAP and HDBSCAN such that the density-based validity index DBCV is maximized across models and layers. Here, for DBCV, the average across clusters is weighted by their respective size such that the noise rate is indirectly included. We re-iterate the effect of the most influential hyperparameters that we tune and state the final value we used:

- **Minimal distance in UMAP**: a low minimal distance in UMAP enhances local cluster density but may also increase noise. We use a value of 0.01 in all experiments.
- **Number of neighbours in UMAP**: the number of neighbors controls the local structure, the smaller the finer it captures local neighborhoods but distorts global structure which is important for concept alignment analysis later. We use a value of 30 in all experiments.
- **Embedding dimensionality in UMAP**: We use the practical limit for HDBSCAN of $F' = 50$ in all experiments.
- **Minimum cluster size in HDBSCAN**: a too small minimum cluster size may identify noise as a cluster, whereas, when too large, distinct clusters will merge. We use a value of 50 in all experiments.
- **Min samples in HDBSCAN**: controls how conservative the algorithm is about noise. We need this to be rather low because of sampling limitations which means that most likely some concept manifolds are not sampled densely enough. We use a value of 20 in all experiments.

Additionally, we assume that clusters are rather uniform in size and select the leaf nodes in the HDBSCAN hierarchical condensed tree as clusters. Sampling one pooled SEQ token or one CLS token from each representation of images within a 25% subset of the ImageNet1-1k train set results in 315.770 feature vectors ϕ_i for clustering. We use the cuML (Raschka et al., 2020) versions of HDBSCAN and UMAP for computation on the GPU.

Cluster label in Concept Atlas To assign a label from the WordNet Hierarchy to each concept cluster, we first assign the ImageNet-1k label of the image from which a token is extracted to its representation feature vector ϕ_i. Then we map this to a label higher in the WordNet hierarchy by the mapping in Tab. 3. We then assign the most frequent label among the cluster members $\{\phi_j^\alpha\}$ to the cluster C^α.

Computation of alignment Our concept-based alignment measure CBA is based on pairs of feature vectors (ϕ_i, ϕ_j). To reduce run-time, we sub-sample 20% of the 315.770 feature vectors before computing CBA.

D Concept Formation Graphs

Notation. Let $C^{l,\alpha}$, with $\alpha \in 1, \dots, N^l$, denote a concept cluster in layer l. For a token $i \in 1, \dots, k$, we use the notation $\boldsymbol{\phi}_i^l \in C^{l,\alpha}$ to indicate that the token representation $\boldsymbol{\phi}_i^l$ in layer l is assigned to that cluster. Let $C^{l,*}$ denote the target concept for which a concept formation graph (CFG) is constructed. Given k tokens sampled from the training dataset and their cluster assignments, the algorithm for constructing the CFG is defined as follows:

1. **Transition matrix calculation:** First, we compute transition matrices $T_{l,l+1} \in \mathbb{Z}^{N^l \times N^{l+1}}$ for each pair of consecutive layers $(l, l+1)$. Each entry represents the count of tokens transitioning from a concept $C^{l,\alpha}$ in layer l to a concept $C^{l+1,\beta}$ in layer $l+1$:

$$\begin{aligned}(T_{l,l+1})_{\alpha\beta} = \#\{&i \in \{1, \dots, k\} \ : \\ &\boldsymbol{\phi}_i^l \in C^{l,\alpha} \text{ and } \boldsymbol{\phi}_i^{l+1} \in C^{l+1,\beta}\}\end{aligned} \tag{12}$$

   where $\#\{\cdot\}$ denotes the count of tokens.

2. **Recursive graph construction:** Initializing the set of CFG nodes with the target concept node $C^{l,*}$, we recursively add all predecessor concept nodes whose "contribution" (the proportion of incoming transitions) surpasses a specified threshold τ. Formally, suppose a concept $C^{l+1,\beta}$ in layer $l+1$ has been added to the CFG. Then, for each concept $C^{l,\alpha}$ in layer l, we include the edge $(C^{l,\alpha}, C^{l+1,\beta})$ and node $C^{l,\alpha}$ to the CFG if:

$$\frac{(T_{l,l+1})_{\alpha\beta}}{\sum_{\gamma=1}^{N^l} (T_{l,l+1})_{\gamma\beta}} > \tau. \tag{13}$$

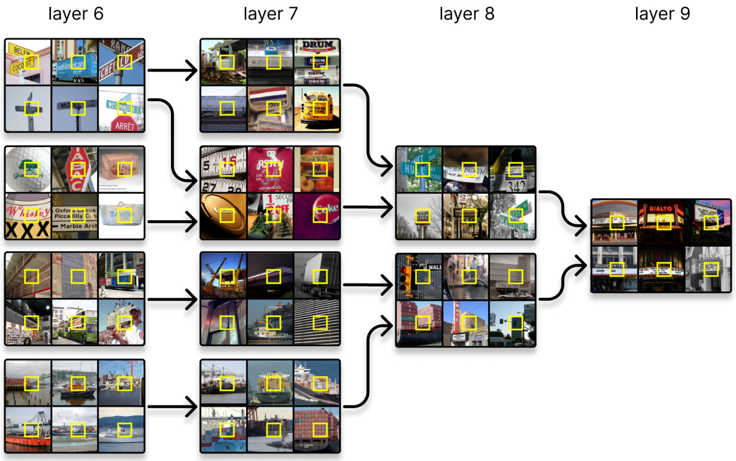

Figure 7: Concept formation graph for a concept in layer 9 of DINO. Each concept is represented by six randomly sampled images containing a token assigned to that concept (highlighted in a yellow frame).

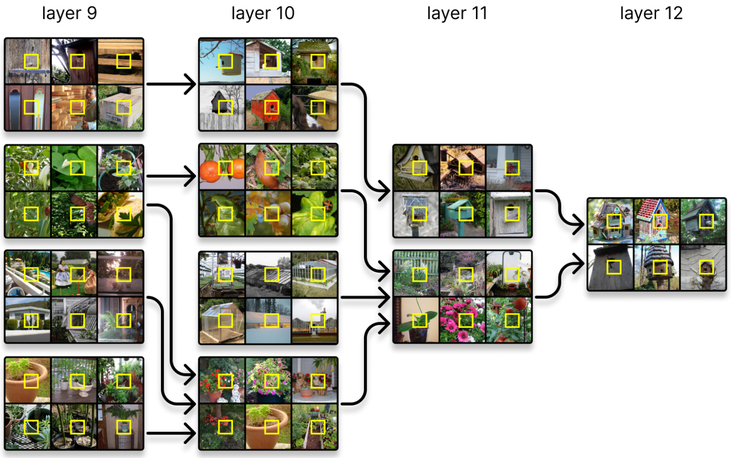

Figure 8: Concept formation graph for a concept in layer 12 of CLIP. Each concept is represented by six randomly sampled images containing a token assigned to that concept (highlighted in a yellow frame).

The resulting CFG is a binary, unidirectional graph. figs. 7 to 9 illustrate additional exemplary CFGs for CLIP and DINO. The CFGs were constructed using the same $k = 315{,}770$ tokens from the ImageNet training dataset that were used for concept discovery, with the threshold parameter set to $\tau = 0.05$. In each image, the “concept” token is highlighted in yellow. The concepts in fig. 3 of the main text are human-labeled.

E Concept alignment analysis

E.1 CLS representations

We investigate concept-based alignment within and across models based on the CLS token representations analogous to the SEQ token analysis in the main paper.

Intra-model alignment First, we focus on the intra-model alignment heatmaps between CLS representations across layers measured by CBA in the upper row of Fig. 10 and compare it to the same analysis between SEQ representations shown in the main paper. For the CLS representations of the FS model we see a very similar pattern as for the SEQ representations. Also for CLIP and MAE, the CLS intra-model alignment mirrors that of the SEQ representations; however, the first two and one

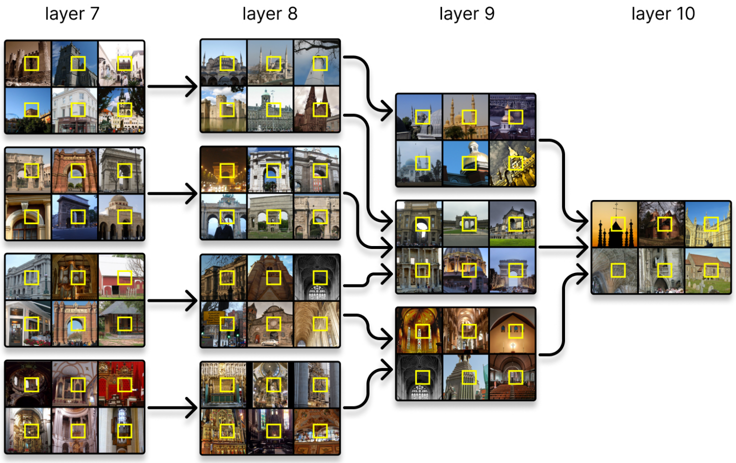

Figure 9: Concept formation graph for a concept in layer 10 of CLIP. Each concept is represented by six randomly sampled images containing a token assigned to that concept (highlighted in a yellow frame).

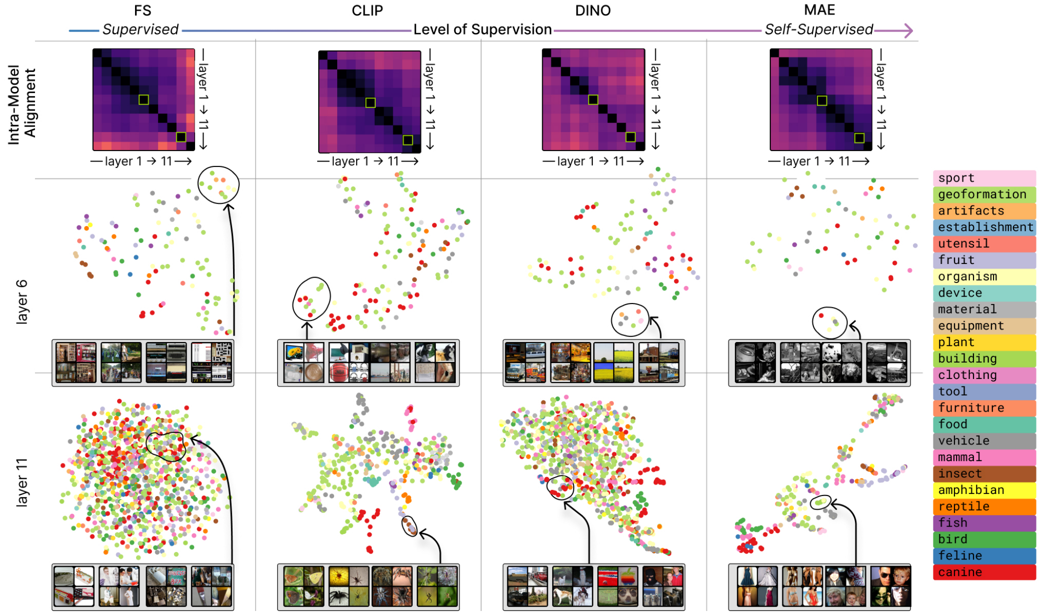

Figure 10: Intra-model relationships based on CLS representations across layers. In the **upper row**, we show CBA to visualize how representations are transformed across layers of the models (darker pixels correspond to higher alignment). In the **center and bottom row** we zoom into the representations at layer 6 and 11 of each model and partition the scalar CBA alignment into single concepts. We show a UMAP embedding constructed from the pairwise distance of concept measured by $d_{cross}(P^\alpha, P^\beta)$. Each point in this *concept atlas* corresponds to a distinct concept P^α. To convey their meaning, we show four random input tokens from the members of the concept cluster C^α marked by a yellow box in the entire image.

blocks, respectively, show significantly lower alignment than in the SEQ tokens. This is reasonable since the model might not use these for processing information in the early blocks. Interestingly, for DINO, the CLS token alignment across layers is significantly lower than the SEQ token alignment. Class label alignment, intrinsic dimensionality of concept clusters and concept count for the CLS representations in Fig. 11 are also similar to the SEQ results except for DINO. Here, DINO CLS concepts exhibit a notable difference to DINO SEQ concepts: the concept count, class alignment,

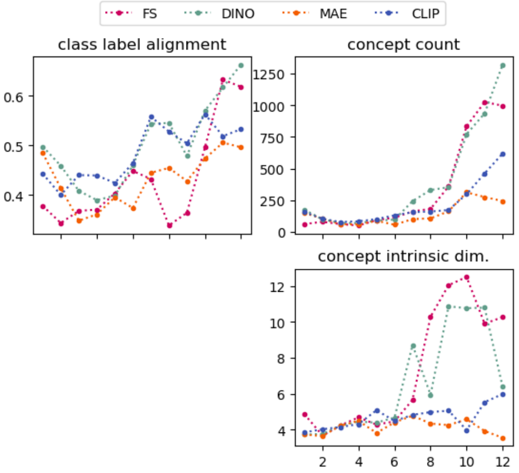

Figure 11: Class label alignment (based on CBA), concept count, and the average intrinsic dimensionality (based on (Facco et al., 2017)) across concepts for CLS representations supplement the intra-model alignment analysis, by providing insights into how well the model aligns with ImageNet-1k labels, the spatial organization of tokens, and the complexity of the learned concepts as they evolve through the layers.

and intrinsic dimensionality increase sharply between blocks 9 and 10 for the CLS representation but not for the SEQ representation. Lastly, the structure of the concept atlases in the lower part of Fig. 10 differs the most from the structure of the SEQ concept atlases for DINO, where CLS concepts at layer 11 are less semantically organized than SEQ concepts. These observations suggest that the differences in how CLS and SEQ tokens represent and abstract information are most pronounced in DINO among the models.

Inter-model alignment Second, we analyze how the CLS representations between two different models differ and present CBA alignment between all layers and model in the upper part of 12. Like for the SEQ representations, CLS representations at layers of the first are more aligned than those of the second half across all models pairs, suggesting that basic foundational features are learned similarly across models, while later layers diverge as the models specialize to concepts serving their pre-training task. However, the overall alignment between models is weaker for CLS representations than for SEQ, also in low layers. Next, we zoom in into the distance $d_{cross}(P^{\alpha}, P^{\beta})$ between concept pairs from CLS representations of DINO layer 3 and MAE layer four in the lower part of 12. We visualize the distance matrix between all concepts and inspect pairs of concepts matched via the Hungarian algorithm. Most of the concepts in the pairs seem to correspond to the color composition of the images. Visual discrepancy is more pronounced for high-distance concepts than for the other pairs with lower distance.

E.2 Additional results for SEQ representations

Intra-model To give a more detailed view of the organization of concepts across the layers of one model, we select the DINO model and show the respective concept atlases at layer one, six and eleven in 13, 14, and 15, respectively. To give an overview of the structure within a concept atlas, we group the concepts in the UMAP embedding via KMeans and show four random concepts for each group. In layer one, many concepts correspond to color, in layer six, they represent mostly textures, and in layer eleven they correspond to abstract concepts. Moslty, concepts within a group are of similar nature.

Inter-model In the main paper, we show fine-grained inter-model concept distances between DINO and MAE at layer six in the center of the both models. Here, we add fine-grained concept distance anaylsis in the first and last part of the models in 16 and 17. We show the full pairwise distance matrix as well as how distances between matched pairs are distributed. We partition the pairs into four regimes of distances with low, medium-low, medium-high, and high distance. The concept pairs between MAE layer 3 and DINO layer 2 seem to correspond mostly to edge detectors

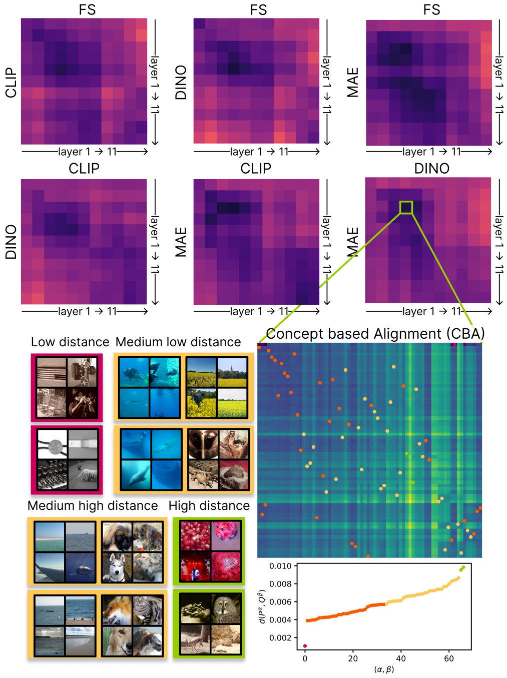

Figure 12: Inter-model relationships based on CLS representations across layers. In the **upper part**, we show CBA to visualize how representations differ between the models (darker pixels correspond to higher alignment). In the **lower part** we zoom in into the concept-wise distances $d_{cross}(P^{\alpha}, P^{\beta})$ between the representation of in MAE layer 2 and DINO layer 3. We give examples of universal concepts with low distance and unique ones with high distance from matches of pairs that minimize the total distance.

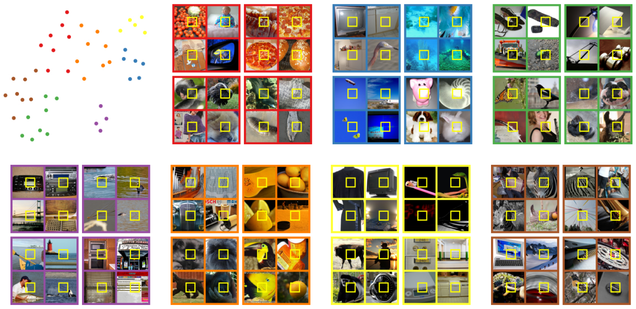

Figure 13: We zoom into the SEQ representations at DINO layer 1 and show a UMAP embedding constructed from the pairwise distance of concepts measured by $d_{cross}(P^{\alpha}, P^{\beta})$. Each point in this *concept atlas* corresponds to a distinct concept P^{α}. To convey their meaning, we show four random input tokens from the members of the concept cluster P^{α}. We dissect the concept atlas into 7 groups and show four random concepts for each group. Concepts representing similar colors lie within the same group, e.g. shades of blue in the blue group or red and orange in the red group.

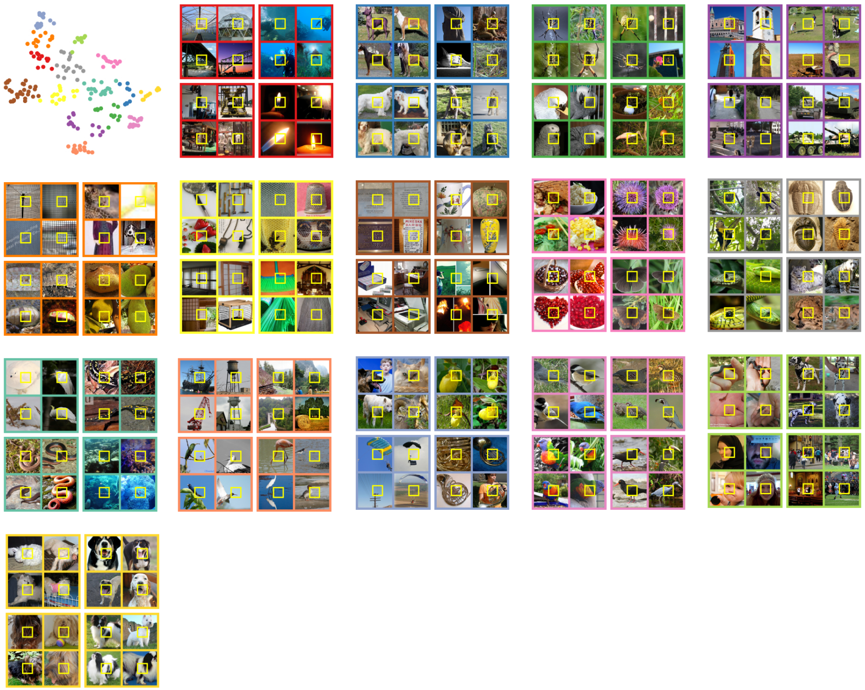

Figure 14: We zoom into the SEQ representations at DINO layer 6 and show a UMAP embedding constructed from the pairwise distance of concepts measured by $d_{cross}(P^{\alpha}, P^{\beta})$. Each point in this *concept atlas* corresponds to a distinct concept P^{α}. To convey their meaning, we show four random input tokens from the members of the concept cluster P^{α}. We dissect the concept atlas into 15 groups and show four random concepts for each group. Most concepts represent a pattern or texture which are similar within each group.

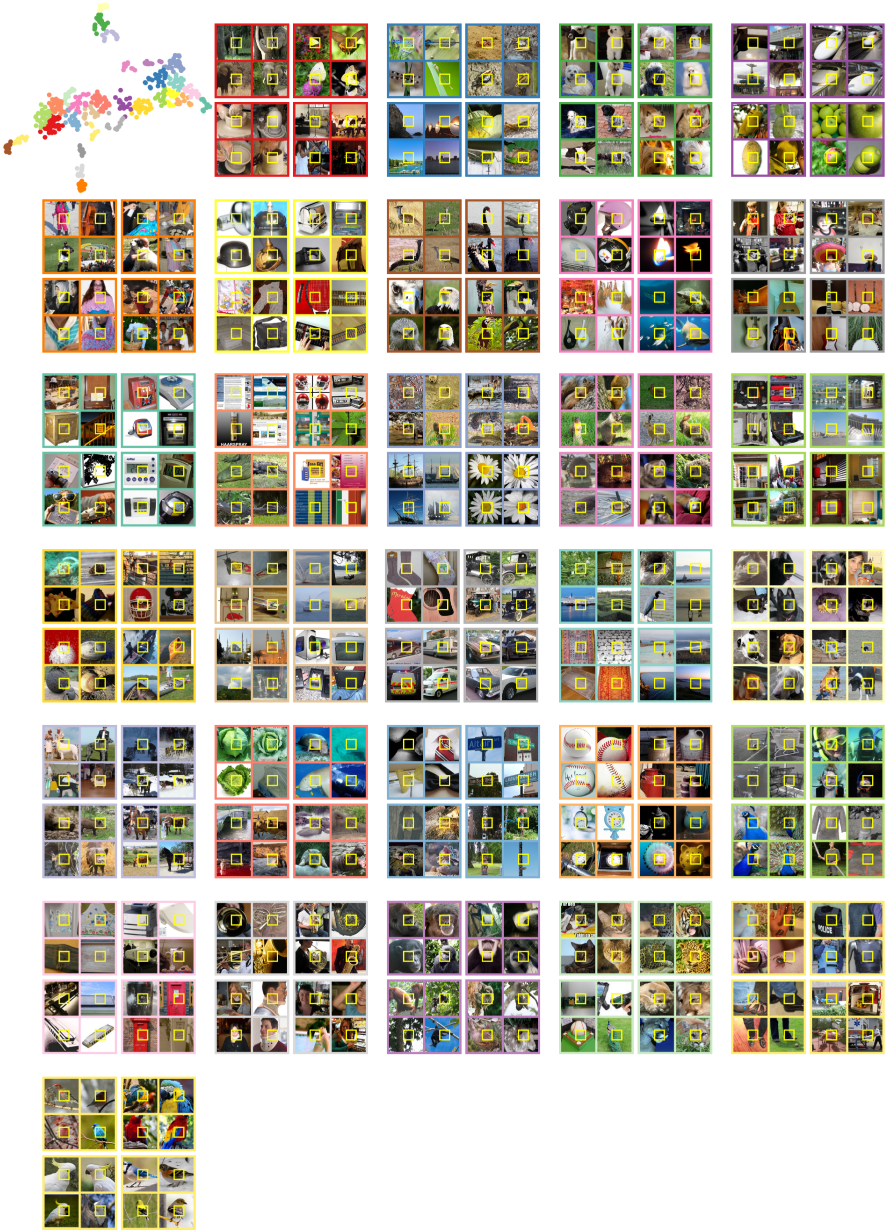

Figure 15: We zoom into the SEQ representations at DINO layer 11 and show a UMAP embedding constructed from the pairwise distance of concepts measured by $d_{cross}(P^\alpha, P^\beta)$. Each point in this *concept atlas* corresponds to a distinct concept P^α. To convey their meaning, we show four random input tokens from the members of the concept cluster P^α. We dissect the concept atlas into 30 groups and show four random concepts for each group. For most groups, these are semantically similar.

or abstract patterns. Among the low and medium-low distance pairs are grid vs. stripes (second low-distance concept pair) and diagonal edge detectors (seond medium-low distance concept pair). The common nature of the medium-high to high-distance pairs is hard to interpret but pairs include warm vs. bright light (first medium-high distance concept pair), and blurriness (third high-distance concept pair). The limitation of visualizing low-level concepts through ImageNet-1k images, as described in the main paper, becomes apparent here. In contrast, the matched concept pairs between MAE and DINO layer 10 are easier to interpret - e.g. owl face or flame.

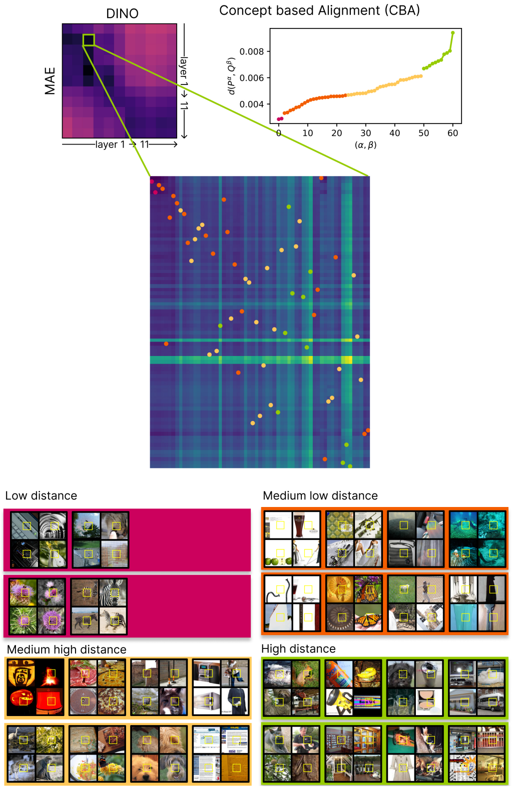

Figure 16: CBA of SEQ concepts across layers of MAE and DINO (darker pixels correspond to higher alignment). We zoom in into the concept-wise distances $d_{cross}(P^\alpha, P^\beta)$ between the representation of layer 2 in MAE and layer 3 in DINO. We give examples of universal concepts with low distance and unique ones with high distance from matches of pairs that minimize the total distance.

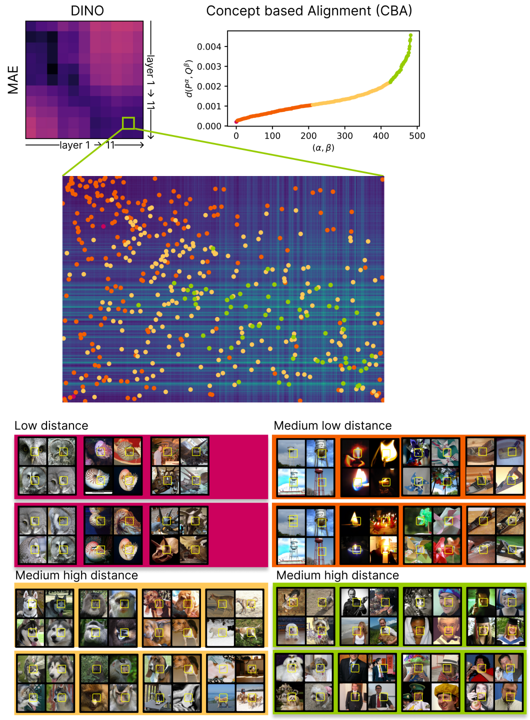

Figure 17: CBA of SEQ concepts across layers of MAE and DINO (darker pixels correspond to higher alignment). We zoom in into the concept-wise distances $d_{cross}(P^{\alpha}, P^{\beta})$ between the representation of layer ten in MAE and DINO. We give examples of universal concepts with low distance and unique ones with high distance from matches of pairs that minimize the total distance.

Category	ImageNet-1k class
amphibian	European fire salamander axolotl bullfrog common newt eft spotted salamander tailed frog tree frog
artifacts	Afghan hound Band Aid Dutch oven Petri dish abacus ashcan backpack ballpoint bannister barrel bath towel bathtub beacon beaker beer bottle beer glass bell cote binder birdhouse book jacket bottlecap brass breakwater breastplate broom bucket cannon carousel carton cassette chain mail chainlink fence chiffonier cleaver cliff dwelling cloak clog cocktail shaker coffee mug comic book cowboy boot crate crib crutch cuirass cup diaper dishwasher dock envelope espresso maker face powder fig fire screen flagpole fountain fountain pen gasmask goblet grasshopper grille grocery store guillotine hair spray hand blower holster honeycomb iron ”jack-o-lantern” joystick ladle lampshade lens cap library lipstick lotion mailbag mailbox manhole cover mask matchstick maze measuring cup megalith menu microwave minibus mixing bowl mobile home mortar mortarboard mosquito net mountain tent muzzle necklace obelisk packet paddle patio pedestal pencil box pencil sharpener perfume pickelhaube picket fence pier piggy bank pill bottle pillow pitcher plastic bag plate rack pole pop bottle pot prayer rug purse quill quilt racket radio rain barrel refrigerator rotisserie rubber eraser running shoe safe saltshaker scabbard school bus schooner scoreboard shield shoji shopping basket shower curtain ski mask sleeping bag sliding door soap dispenser soup bowl space bar spotlight steel arch bridge stone wall stove street sign stretcher sunscreen suspension bridge swab swing teddy television thatch theater curtain thimble tile roof totem pole traffic light tray triumphal arch trolleybus tub turnstile umbrella vacuum vase viaduct waffle iron washbasin washer water bottle water jug water tower web site whiskey jug window screen window shade wine bottle worm fence wreck yurt
bird	African grey American coot American egret European gallinule albatross bald eagle bee eater bittern black grouse black stork black swan brambling bulbul bustard chickadee cock coucal crane dowitcher drake flamingo goldfinch goose great grey owl hen hornbill house finch hummingbird indigo bunting jacamar jay junco king penguin kite limpkin little blue heron lorikeet macaw magpie ostrich oystercatcher partridge pelican prairie chicken ptarmigan quail red-backed sandpiper red-breasted merganser redshank robin ruddy turnstone ruffed grouse spoonbill sulphur-crested cockatoo toucan vulture water ouzel white stork
building	apiary barn boathouse castle church cinema greenhouse home theater monastery mosque palace planetarium prison restaurant stage stupa vault
canine	African hunting dog Airedale American Staffordshire terrier Appenzeller Arctic fox Australian terrier Bedlington terrier Bernese mountain dog Blenheim spaniel Border collie Border terrier Boston bull Bouvier des Flandres Brabancon griffon Brittany spaniel Cardigan Chesapeake Bay retriever Chihuahua Dandie Dinmont Doberman English foxhound English setter English springer EntleBucher Eskimo dog French bulldog German short-haired pointer Gordon setter Great Dane Great Pyrenees Greater Swiss Mountain dog Ibizan hound Irish setter Irish terrier Irish water spaniel Irish wolfhound Italian greyhound Japanese spaniel Kerry blue terrier Labrador retriever Lakeland terrier Leonberg Lhasa Maltese dog Mexican hairless Newfoundland Norfolk terrier Norwegian elkhound Norwich terrier Pekinese Pembroke Pomeranian Rhodesian ridgeback Rottweiler Saint Bernard Saluki Samoyed Scotch terrier Scottish deerhound Sealyham terrier Shetland sheepdog Shih-Tzu Siberian husky Staffordshire bullterrier Sussex spaniel Tibetan mastiff Tibetan terrier Walker hound Weimaraner Welsh springer spaniel West Highland white terrier Yorkshire terrier affenpinscher basenji basset beagle black-and-tan coonhound bloodhound bluetick borzoi briard bull mastiff cairn chow clumber cocker spaniel collie coyote curly-coated retriever dalmatian dhole dingo flat-coated retriever giant schnauzer golden retriever grey fox groenendael hyena keeshond kelpie kit fox komondor kuvasz malamute malinois miniature pinscher miniature poodle miniature schnauzer otterhound papillon pug red fox red wolf redbone schipperke silky terrier soft-coated wheaten terrier standard poodle standard schnauzer timber wolf toy poodle toy terrier vizsla whippet white wolf wire-haired fox terrier
clothing	Christmas stocking Loafer Old English sheepdog Windsor tie abaya academic gown apron bathing cap bearskin bib bikini bolo tie bonnet bow tie brassiere bulletproof vest cardigan chest cowboy hat crash helmet dishrag feather boa fur coat gown handkerchief hook hoopskirt jean jersey kimono knee pad lab coat maillot military uniform miniskirt mitten overskirt pajama paper towel poncho sandal sarong seat belt shower cap sock sombrero stole suit sweatshirt swimming trunks trench coat velvet vestment wallet wig wool
device	accordion acoustic guitar analog clock assault rifle banjo barometer bassoon binoculars bow buckle bullet train candle car mirror car wheel cash machine cello chime combination lock desktop computer digital clock digital watch disk brake drum drumstick electric fan electric guitar flute gas pump gong hair slide hammer hamper hand-held computer hard disc harmonica harp hatchet horn hourglass laptop lighter loudspeaker loupe magnetic compass maraca marimba maypole microphone missile monitor mouse mousetrap neck brace notebook oboe odometer oil filter organ oxygen mask paddlewheel padlock paintbrush panpipe parking meter pick ”potters wheel” projector puck radiator radio telescope remote control revolver rifle safety pin sax scale screen sewing machine ski slide rule slot slug snorkel solar dish space heater spider web steel drum stethoscope stopwatch strainer sundial sunglass sunglasses switch syringe thresher toaster torch tripod trombone typewriter keyboard upright vending machine violin wall clock whistle
equipment	CD player Polaroid camera balance beam barbell ”carpenters kit” cassette player cellular telephone computer keyboard croquet ball crossword puzzle dial telephone drilling platform dumbbell golf ball golfcart horizontal bar iPod jigsaw puzzle modem oscilloscope parachute parallel bars pay-phone photocopier ping-pong ball plate punching bag reel reflex camera soccer ball tape player
establishment	bakery barbershop bookshop butcher shop confectionery shoe shop tobacco shop toyshop
feline	Egyptian cat Persian cat Siamese cat catamount cheetah coil cougar leopard lion panther snow leopard tabby tiger tiger cat
fish	anemone fish barracouta coho eel electric ray gar goldfish great white shark hammerhead lionfish puffer rock beauty stingray sturgeon tench tiger shark
food	French loaf bagel burrito carbonara cheeseburger chocolate sauce consomme cucumber dough eggnog espresso guacamole hay hot pot hotdog ice cream ice lolly mashed potato meat loaf pizza potpie pretzel red wine trifle
fruit	Granny Smith acorn buckeye hip jackfruit lemon orange pineapple rapeseed strawberry
furniture	altar barber chair bassinet beaver bookcase china cabinet cradle desk dining table entertainment center file folding chair four-poster medicine chest milk can mink otter park bench pool table rocking chair studio couch table lamp throne toilet seat wardrobe
geological formation	alp bubble cliff coral reef dome geyser lakeside promontory sandbar seashore valley volcano
insect	ant bee cabbage butterfly cicada cricket damselfly dragonfly dung beetle fly ground beetle lacewing ladybug leaf beetle leafhopper long-horned beetle lycaenid mantis monarch peacock rhinoceros beetle ringlet sulphur butterfly tiger beetle walking stick weevil
mammal	African elephant American black bear Angora Arabian camel Indian elephant Madagascar cat Sus scrofa armadillo baboon bighorn bison black-footed ferret brown bear capuchin chimpanzee colobus dugong echidna fitch fox squirrel gazelle gibbon gorilla grey whale guenon guinea pig hamster hare hartebeest hippopotamus ibex ice bear impala indri killer whale koala langur lesser panda llama macaque marmoset marmot meerkat mongoose orangutan ox panda patas platypus polecat porcupine proboscis monkey ram sea lion siamang sloth bear spider monkey squirrel monkey three-toed sloth titi tusker wallaby warthog water buffalo wombat wood rabbit zebra
material	chain cornet doormat groom knot spindle toilet tissue
musical	grand piano
organism	American lobster Dungeness crab German shepherd admiral agaric badger ballplayer barn spider black and gold garden spider black widow bolete boxer brain coral centipede chiton cockroach conch coral fungus crawfish dam ear earthstar fiddler crab flatworm garden spider gyromitra harvester harvestman hen-of-the-woods hermit crab hog howler monkey isopod jellyfish king crab mushroom nematode nipple printer rock crab rule scorpion scuba diver sea cucumber sea slug sea urchin snail spiny lobster starfish stinkhorn tarantula tick trilobite weasel wing wolf spider
plant	acorn squash artichoke banana bell pepper broccoli butternut squash cardoon cauliflower corn custard apple daisy head cabbage ocarina pinwheel pomegranate sea anemone sorrel spaghetti squash ”yellow ladys slipper” zucchini
reptile	African chameleon African crocodile American alligator American chameleon Gila monster Indian cobra Komodo dragon agama alligator lizard banded gecko boa constrictor box turtle common iguana diamondback frilled lizard grass snake green lizard green mamba green snake hognose snake king snake leatherback turtle loggerhead mud turtle night snake ringneck snake rock python sand viper sea snake sidewinder terrapin thunder snake triceratops vine snake water snake whiptail
sport	baseball basketball football helmet rugby ball tennis ball volleyball
tool	can opener chain saw corkscrew lawn mower letter opener lumbermill nail plane plow plunger power drill screw screwdriver shovel
utensil	Crock Pot caldron coffeepot frying pan spatula teapot wok wooden spoon
vehicle	Model T aircraft carrier airliner airship ambulance amphibian balloon barrow beach wagon bicycle-built-for-two bobsled cab canoe catamaran chambered nautilus container ship convertible dogsled electric locomotive fire engine fireboat forklift freight car garbage truck go-kart gondola half track horse cart jeep jinrikisha lifeboat limousine liner minivan moped motor scooter mountain bike moving van oxcart passenger car pickup pirate racer recreational vehicle shopping cart snowmobile snowplow space shuttle speedboat sports car steam locomotive streetcar submarine tank tow truck tractor trailer truck tricycle trimaran unicycle wagon warplane yawl

Table 3: Mapping between categories from the WordNet Hierarchy and the ImageNet-1k classes used for assigning a category to the concept clusters.