Brain Decodes Deep Nets

Huzheng Yang **James Gee*** **Jianbo Shi***
University of Pennsylvania
https://huzeyann.github.io/brain-decodes-deep-nets

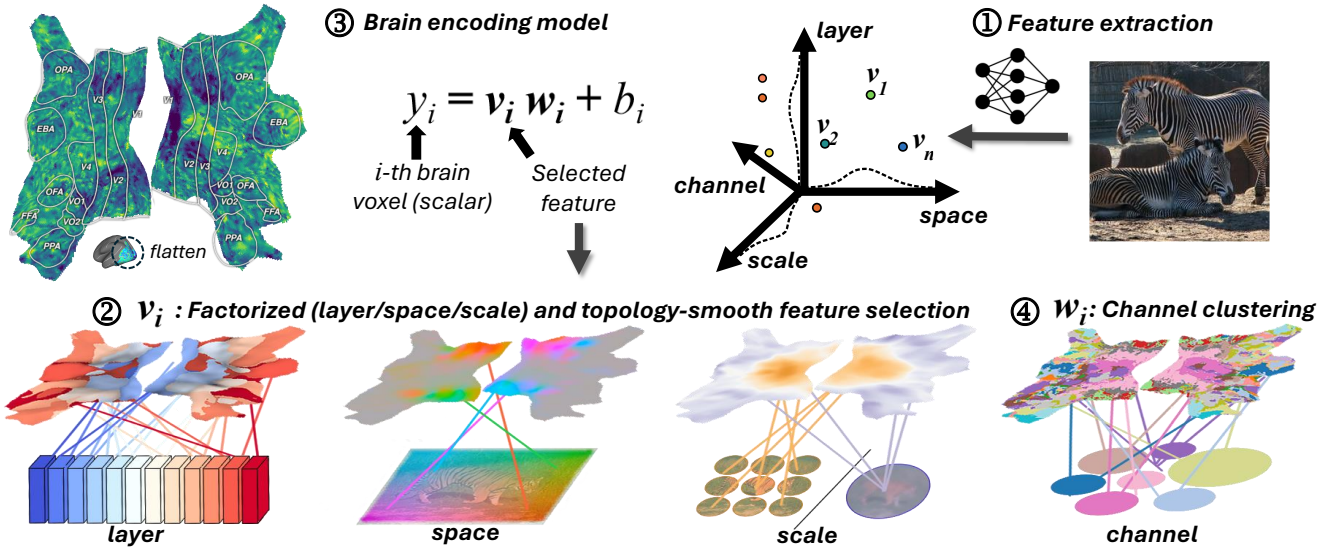

Figure 1. **Visualize Deep Networks in the Brain**. The training objective of the brain encoding model is to predict the brain's fMRI signal in response to an image stimulus. 3D visual brain surface is flattened into 2D for better visualization. ① Image features are extracted from a pre-trained network. ② Feature selection for each voxel is randomly initialized and learned using the brain encoding training objective. The selection is ***factorized*** in the layer/space/scale axis; the ***topological constraint*** improves selection smoothness and confidence. ③ Linearized brain encoding model. ④ After training, linear weights are used to cluster channels. We use the resulting brain-to-network mapping together with the known knowledge of the brain to answer the question *"how do deep networks work?"*.

Abstract

We developed a tool for visualizing and analyzing large pre-trained vision models by mapping them onto the brain, thus exposing their hidden inside. Our innovation arises from a surprising usage of brain encoding: predicting brain fMRI measurements in response to images. We report two findings. First, explicit mapping between the brain and deep-network features across dimensions of space, layers, scales, and channels is crucial. This mapping method, FactorTopy, is plug-and-play for any deep-network; with it, one can paint a picture of the network onto the brain (literally!). Second, our visualization shows how different training methods matter: they lead to remarkable differences in hierarchical organization and scaling behavior, growing with more data or network capacity. It also provides insight into fine-tuning: how pre-trained models change when adapting to small datasets. We found brain-like hierarchically organized network suffer less from catastrophic forgetting after fine-tuned.

1. Introduction

The brain is massive, and its enormous size hides within it a mystery: how it efficiently organizes many specialized modules with distributed representation and control. One clue it offers is its feed-forward hierarchical organization (Figure 2). This hierarchical structure facilitates efficient computation, continuous learning, and adaptation to dynamic tasks.

Deep networks are enormous, containing billions of parameters. Performances keep improving with more training data and larger size. It doesn't seem to matter if the network is trained under the supervision of labels, weakly supervised with image captions, or even self-supervised without human-provided guidance. Its sheer size also hides another mystery: as its size increases, it can be fine-tuned successfully to many unseen tasks.

*: Equal advising.

What can these two massive systems, the brain and deep network, tell about each other? By identifying 'what' deep features are most relevant for each brain voxel fMRI prediction, we can obtain a picture of deep features mapped onto a brain (literally), as shown by the brain-to-network mapping in Figure 1.

The key insight is that deep networks trained with the same architecture, but different objectives and data, produce drastically different computation layouts of intermediate layers, even if they can produce similar brain encoding scores and other downstream task scores. For example, we found intermediate layers of CLIP align hierarchically to the visual brain. However, there are unexpected non-hierarchical bottom-up and top-down structure in supervised classification and segmentation-trained models. Moreover, for many models, when scaling up in parameters and training data, they tend to lose hierarchical alignment to the brain, except CLIP, which improved hierarchical alignment to the brain after scaling up.

Suppose the brain's hierarchical organization is a template for efficient, modular, and generalizable computation; an ideal computer vision model should align with the brain: the first layer of the deep network matches the early visual cortex, and the last layer best matches high-level regions. Our fine-tuning results show that networks with more hierarchy organization tend to (qualitatively) maintain their hidden layers better after fine-tuning on small datasets, thus suffering less (quantitatively) from catastrophic forgetting. We conjecture that better alignment to the brain is one way to find a robust model that adapts to dynamic tasks and scales better with larger models and more data.

Our analysis crucially depends on a robust mapping between deep 4D features: spatial, layer, channel, and scale (class token vs local token) to the brain. Our fundamental assumption is that this mapping should be: a) *brain-topology constrained*, and b) *factorized* in feature dimensions of space, layer, channel, and scale. This is important because independent 4D image features to brain mapping are highly unconstrained, and learning a shared mapping across images, with brain-topology constraint and factorized representation, is statistically more stable.

Our contribution is summarized as the following:

1. We introduce a *factorized, brain-topological smooth* selection that produces an explicit mapping between deep features: space, layer, channel, and scale (class token vs local token) to the brain.
2. We pioneer a new network visualization by coloring the brain using layer-selectors, exposing the inner workings of the network.
3. We found that brain-like hierarchically organized networks suffer less from catastrophic forgetting after fine-tuning.

[1]: The Algonauts 2023 competition: http://algonauts.csail.mit.edu/

2. Background and Related Work

Hierarchy of the Visual Brain In Figure 2, visual brain is organized into regions, each region has specialized functions. Image processing in visual brain is organized in a hierarchical and feed-forward fashion. Starting from region V1 to V4, neurons were found to have increasing receptive field size and represent more abstract concepts [12, 13, 61], the late visual brain has semantic regions such as face (FFA), body (EBA), and place (OPA, PPA).

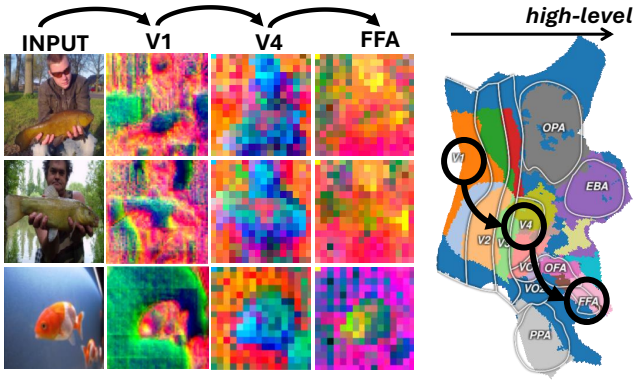

Figure 2. **Image features (selected channels) for brain ROIs**. V1 is orientation filtering, V4 segmentation, FFA face-selective.

Brain Encoding Benchmarks Open challenge and competitions on brain encoding model have generated broad interests [7, 8, 17, 48, 49, 56, 59]. Large-scale open-source datasets are growing rapidly in both quantity and quality [1, 5, 16, 22, 29]. The Algonauts[1] 2023 competition [17] is the first to use a massive high-quality 7-Tesla fMRI dataset [1]. The high-quality and large-scale of this datasets enabled models that can recover brain-to-space mapping from naturalistic image stimuli [44], which was only possible with synthetic stimuli [13]. Our brain encoding model methods is a direct extension of the Algonauts 2023 competition winning method *Memory Encoding Model* [63]. In this work, we added a scale axis for feature selection.

Explain Brain by Deep Networks After fitting brain encoding models to predict brain response, gradient-based methods have been used to explain how brain works: orientation-selective neurons in V1 [14, 35, 44], category-selective regions in late visual brain [26, 32, 33, 39, 42, 47]. Gradient-based methods can also generate maximum-excited images [3, 18, 28, 52, 62]. Meanwhile, studies try to find the best performance pre-trained model for each brain ROI [10, 37, 46, 58, 66] from a zoo of supervised [25, 27, 41, 51], self-supervised [6, 19, 21, 30, 38], image generation [43], and 3D [36, 45, 55] models. Features can be efficiently cached and are plug-in-and-play [20, 54, 57]. Different from the main-stream study that use deep networks to explain the brain's functionality. In this work, we use existing knowledge of the brain's functionality to explain feature computation in deep networks.

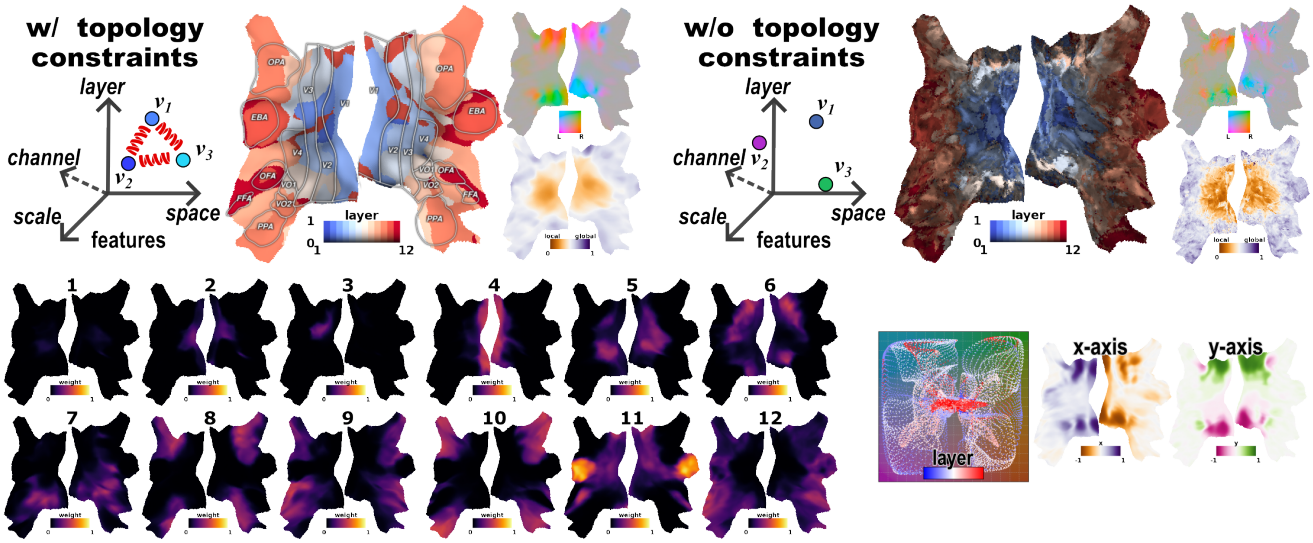

Figure 3. **Topological Constrained, Factorized, Brain-to-Network Selectors for CLIP**. ***Top***: factorized-selectors trained with topological constraints improved confidence of the mapping (color brightness) and mapping smoothness (colored as Section 3.2). ***Bottom left***: individual layer-selector weight $\hat{\boldsymbol{\omega}}^{layer}$, note layer 4 is mostly aligned with V1, and the last two are aligned with the body (EBA) and face (FFA) region. ***Bottom right***: space-selector $\hat{\boldsymbol{u}}^{space}$: 3D voxels, dots, are mapped to the image space with color dots indicating the layers. For later layers, only center image regions are selected.

3. Methods: Brain Encoding Model

Figure 1 presents an overview of our methods. In the brain encoding task, one needs to predict a large number voxels (vertices), of visual cortex's fMRI responses as a function of the observed image. This encoding task is underconstrained: since each subject has her/his unique mental process, a successful brain encoding model needs to be highly individualized, thus significantly reducing the training example per voxel. Most of the current approaches treat each brain voxel independently. This leads to a major reduction in signal-to-noise ratio, particularly for our analysis.

Our fundamental innovations are two-fold. First, we enforce brain-and-network ***topology-constrained*** prediction. Brain voxels are not independent but are organized locally into similar "tasks", and globally into diverse functional regions. Similarly, Neural networks show local feature similarity across adjacent layers while ensuring diversity for faraway ones. The local smoothness constraints significantly reduce uncertainties in network-to-brain mapping.

Second, we propose a ***factorized*** feature selection across three independent dimensions of space, layers, and scales (local vs global token). This factorized representation leads to a more robust estimation because feature selection in each dimension is more straightforward, and learning can be more efficient across training samples. For example, the spatial feature selection only needs to find the center of the pixel region for each brain voxel, similar to retinotopy. The layer or scale selection estimates the size of the pixel region: the early layer typically has a smaller receptive field size. Note that the factorized feature selection is *soft*: multiple layers or spatial locations can be selected, as determined by the brain prediction training target.

3.1. Factorized, Topological Smooth, Brain-to-network Selection (FactorTopy)

We used a pre-trained image backbone model (ViT) to process input image $\boldsymbol{X}$ into features $\boldsymbol{V}$. The entire feature $\boldsymbol{V}$ is organized along four dimensions: space, layer, scale (class token and local tokens), and channels.

The current state-of-the-art methods [1] compute a layer-specific, scale-specific, 2D spatial feature selection mask to pool features $\boldsymbol{V} \in \mathbb{R}^{L\times C\times H\times W}$ along spatial dimension $H \times W$ into a vector of $\mathbb{R}^{L\times C}$, where L denotes layer and C is channel. Instead, we propose a ***factorized*** feature selection method where, for each voxel, we select the corresponding space, layer, and scale in each dimension.

Essentially, a voxel asks: 'What is the best x-factor for my brain prediction?' where the x-factor is one of the layer, space, scale, or channel dimensions.

1) space selector. $selSpace: \ \mathbb{R}^{N\times 3} \rightarrow \hat{\boldsymbol{u}}^{space} \in \mathbb{R}^{N\times 2}$, maps brain voxels' 3D coordinates into 2D image coordinates, where N is number of voxels. We used linear interpolation `Interp` to extract $\bar{\boldsymbol{\nu}}^{local}_{i,l} \in \mathbb{R}^{1\times C}$.

2) layer selector. $selLayer: \ \mathbb{R}^{N\times 3} \rightarrow \hat{\boldsymbol{\omega}}^{layer} \in \mathbb{R}^{N\times L}$, produces $\hat{\omega}^{layer}_{i,l} \in [0,1]$ weight for each layer l, such that $\sum_{l=1}^{L} \hat{\omega}^{layer}_{i,l} = 1$. We take a weighted channel-wise average of feature vectors $\bar{\boldsymbol{\nu}}_{i,l}$ across all layers.

3) scale selector: $selScale: \ \mathbb{R}^{N\times 3} \rightarrow \hat{\boldsymbol{\alpha}}^{scale} \in \mathbb{R}^{N\times 1}$, computes a scalar $\hat{\alpha}^{scale}_{i} \in [0,1]$ as the weight for local $\bar{\boldsymbol{\nu}}^{local}_{i,l}$ vs global token $\bar{\boldsymbol{\nu}}^{global}_{*,l}$. Note that $\bar{\boldsymbol{\nu}}^{local}_{i,l}$ is unique for each voxel, $\bar{\boldsymbol{\nu}}^{global}_{*,l}$ is same for all voxels.

Taking weighted averages over channels across layers could be problematic because channels in each layer represent different information. We need to preemptively align

the channels into a shared D dimension channel space. Let B_l be a layer-unique channel transformation:
channel align. $B_l(\boldsymbol{V}_l): \mathbb{R}^{C\times M} \rightarrow \mathbb{R}^{D\times M}$, where $M = (H \times W + 1)$.

The brain encoding prediction target $\boldsymbol{Y} \in \mathbb{R}^{N\times 1}$ is beta weights (amplitude) of hemodynamic response (pulse) function [40]. Denote scalar y_i the individual voxel $i \in \{1, 2, \ldots, N\}$ response. To obtain the final brain prediction scalar y_i, we apply feature selection across the channels:
4) channel selector. $\boldsymbol{w}_i: \mathbb{R}^D \rightarrow \mathbb{R}^1$, where $\boldsymbol{w}_i$ answers, 'Which is the best channel for predicting this brain voxel?' Putting it all together, we have

$$\begin{aligned} \boldsymbol{V} &= \texttt{ViT}(\boldsymbol{X}) \\ \bar{\boldsymbol{\nu}}^{local}_{i,l} &= \texttt{Interp}(\hat{\boldsymbol{u}}^{space}_i; B_l(\boldsymbol{V}_l)) \\ \boldsymbol{v}_i &= \sum_{l=1}^{L} \hat{\omega}^{layer}_{i,l}((1-\hat{\alpha}^{scale}_i)\bar{\boldsymbol{\nu}}^{local}_{i,l} + \hat{\alpha}^{scale}_i \bar{\boldsymbol{\nu}}^{global}_{*,l}) \\ y_i &= \boldsymbol{v}_i \boldsymbol{w}_i + b_i \end{aligned} \tag{1}$$

Topological Smooth. The factorized selector explicitly maps the brain and the network. The topological structure of the corresponding brain voxels should also constrain this mapping. The smoothness constraint can be formulated as Lipschitz continuity [2]: nearby brain voxels should have similar space, layer, and scale selection values. We apply sinusoidal position encoding [36] to brain voxel.

3.2. Visualization and Coloring

To visualize layer-to-brain mapping, we assign each voxel a color cue value associated with the layer with the highest layer selection value: $\texttt{argmax}_L(\hat{\boldsymbol{\omega}}^{layer}) \in \mathbb{R}^{N\times 1}$. We assign voxel color brightness with a confidence measure $\boldsymbol{s} \in \mathbb{R}^{N\times 1}$ of $\hat{\boldsymbol{\omega}}^{layer}$:

$$s_i = 1 - \frac{\sum_{l=1}^{L} \hat{\omega}^{layer}_{i,l} \log \hat{\omega}^{layer}_{i,l}}{\sum_{l=1}^{L} \frac{1}{L} \log \frac{1}{L}} \tag{2}$$

Note that s_i equals 1 when $\hat{\boldsymbol{\omega}}^{layer}_i$ is a one-hot vector, and 0 when it is uniform. In Figure 3, we compare layer-selector trained with vs. without topological smooth constraints using this layer-to-brain color scheme. Topological smoothness significantly improved selection certainty.

4. Results

For a fair comparison, we keep the same ViT network architecture while varying how the network is trained and the dataset used (Table 2). In Fig. 5, we display network layer-to-brain mapping for several popular pre-trained models. An overview of our experiments:

1. *What can relative brain prediction scores tell us?*
2. *How do supervised and un-supervised training objectives change brain-network alignment?*
3. *Do more data and larger model sizes lead to a more evident hierarchical structure?*
4. *What happens to a pre-trained network when fine-tuning to a new task with small samples?*
5. *Can the network channels be grouped to match well with brain functional units?*

Dataset We used Nature Scenes Dataset (NSD) [1] for this study. Briefly, NSD provides 7T fMRI scan when watching COCO images [31]. A total of 8 subjects each viewed 3 repetitions of 10,000 images. We used the preprocessed and denoised single-trail data of the first 3 subjects [40]. We split 27,750 public trials into train validation and test sets (8:1:1) and ensured no data-leak of repeated trials.

4.1. Brain Score for Downstream Tasks Prediction

The key finding is that a network with a high prediction score on a specific brain region is better suited for a relevant downstream task. CLIP, DiNOv2 and Stable Diffusion have overall high performance.

Let $R^2 = 1 - \sum(y_{i,m} - \hat{y}_{i,m})^2 / \sum(y_{i,m} - \bar{y}_{i,m})^2$ be the brain score metrics, R^2 is computed for each voxel i over the test-set m. We report the raw score without dividing by noise ceiling or averaging repeated trials [17]. We compared the brain score for each model to the 'max' model constructed by model-wise maximum for each voxel. We show the raw R^2 in Figure 4, and ROI-wise root sum squared difference to the 'max' in Table 1.

Model	Dataset	Root Sum Squared Difference $R^2 \downarrow$					
		V1	V2V3	OPA	EBA	FFA	PPA
Known Selectivity		orientation		navigate	body	face	scene
max		0.237	0.215	0.097	0.185	0.186	0.134
CLIP [41]	DC-1B [15]	***0.032***	**0.023**	***0.011***	***0.015***	**0.005**	**0.006**
DiNOv2 [38]	LVD-142M	0.033	0.026	0.021	**0.013**	***0.008***	***0.007***
SAM [27]	SA-1B	0.037	0.033	0.025	0.065	0.056	0.033
MAE [21]	IN-1K	**0.031**	***0.025***	**0.008**	0.029	0.017	0.009
MoCov3 [6]	IN-1K	***0.032***	0.027	0.014	0.031	0.015	0.011
ImageNet [11]	IN-1K	0.037	0.032	0.024	0.028	0.019	0.015
SD (T20) [43]	LAION-5B [50]	0.047	0.050	0.029	0.056	0.052	0.032
SD (T40) [43]	LAION-5B	**0.031**	0.030	0.021	0.018	0.019	0.013

Table 1. **Brain Score**. Raw R^2 for max of all models and root sum squared difference for other models.

Figure 4 shows that the fovea regions of early visual cortex are highly predictable, and so are higher regions of EBA and FFA, followed by PPA. In Table 1, we found DiNOv2 and CLIP predict well on EBA and FFA but poorly for early visual regions; MAE and SAM are the opposite. Stable Diffusion (SD) features, described in the next section, perform well in all regions. This finding is consistent with recent works that show SD features are helpful for many visual tasks, from segmentation to semantic correspondence [53, 60]. It could also explain why a combination of DiNOv2 for coarser semantic correspondence with SD for finer alignment could work well [57, 64].

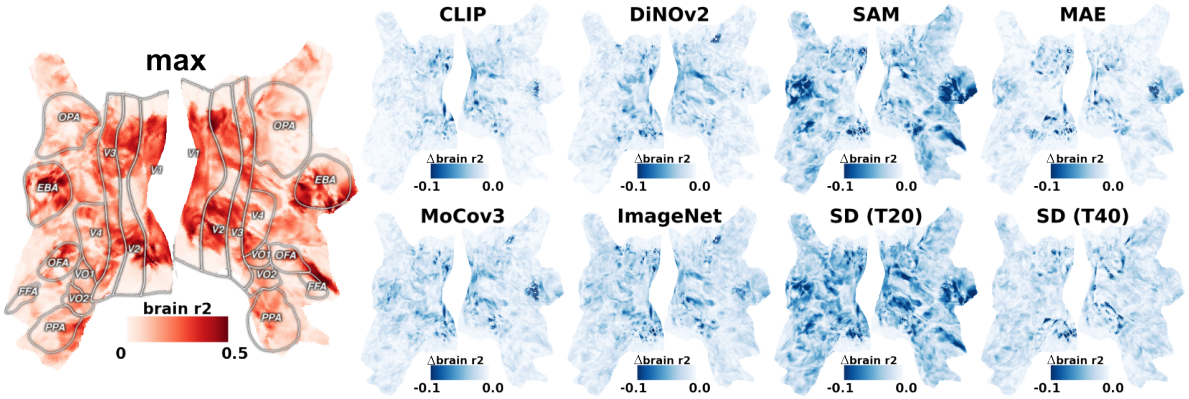

Figure 4. **Brain Score**. ***Left***: raw brain score R^2. ***Right***: difference of score to the model-wise max score (left). ***Insights***: **1)** CLIP and DiNOv2 predict semantic regions better but relatively weak for early visual, **2)** SAM and MAE are better at early visual region but weaker for body (EBA) and face (FFA) region, **3)** Stable Diffusion (SD) shows a good prediction in all regions overall.

4.2. Training Objectives and Brain-Net Alignment

The key finding is training objective matters: **1)** supervised methods show a more detailed delineation of network-to-brain mapping compared to self-supervised ones; **2)** ImageNet and SAM show the last layer mapped to the middle region of the brain; **3)** Stable Diffusion features show more detailed delineation between the time steps than between the UNet encoder or decoder layers.

The layer multi-selector output indicates, "*within one model, which layer best predicts this brain region?*". Even though the mapping differs for subjects (Figure 5), the pattern of subject difference is consistent in both CLIP and ImageNet models: subject #3 had considerably low confidence in early visual brain, and subjects #2 and #3 are missing the FFA (face) region that subject #1 has.

For supervised pre-trained models, Figure 5 shows CLIP's [41] last layer is close to EBA for subject average and EBA/FFA for subject #1, probably because the training data contain languages related to body and face. Surprisingly, ImageNet's [11] last layer is close to the mid-level lateral stream, suggesting that simple image labels are more primitive than text language. SAM's [27] final layer is close to the mid-level ventral and parietal stream, indicating segmentation as a mid-level visual task. These observations suggest a bottom-up feature computation and top-down task prediction in ImageNet and SAM.

For the self-supervised models, the final layer of DiNOv2 (DiNOv1+iBOT) [4, 38, 65] and MAE [21] is missing from the network-to-brain mapping, which indicates the last stage of un-supervised mask reconstruction deviates from the brain tasks. For MoCov3 [6], there's a trend that the second-last layer matched more with the ventral stream ("what" part of the brain) than the parietal stream ("where" part), indicating self-contrastive learning is more focused on the semantics rather than spatial relationship [57].

We also analyzed Stable Diffusion [43] by 1) fixing the time step and selecting layers, and 2) fixing the UNet decoder layer 6 and selecting time steps. We followed the "inversion" [34] time steps feature extraction and used a total of T=50 time steps. In Figure 6, layer selection showed that the diffusion model has less separation for early and late regions; this was true for both T=25, T=40, encoders and decoders. Time step selection showed diffusion model early time steps (T<25) deviate from the brain tasks. The confidence (Section 3.2) of time step selection was relatively high for EBA at (T=30) and for mid-level visual stream at (T=35, T=40). Overall, our results indicate that 1) the diffusion model has less feature separation across layers but instead is separated across time steps, and 2) global features are more in the middle-time steps, while local features are more aligned with the mid-to-late time steps.

4.3. Network Hierarchy and Model Sizes

The key findings are: **1)** CLIP shows a more substantial alignment of hierarchical organization with the brain; **2)** when scaling with more data and bigger model size, CLIP shows an improvement in its brain-hierarchical alignment, while others show a decrease.

We propose a measure called ***hierarchy slope*** by putting predefined brain ROI regions into a number-ordering and fitting a linear regression as a function of their layer selector output $\hat{\omega}^{layer}$. We used only coarse brain regions and did not consider feedback computation in the brain.

Hierarchy slope Let $\hat{\iota}_i = \sum_{l=1}^{L} \frac{l-1}{L-1} \hat{\omega}_{i,l}^{layer}$ be a scalar that represents vector layer selector weights $\hat{\omega}_i^{layer}$, such that $\hat{\iota}_i \in [0, 1]$. We pre-defined a four-level brain structure: 1) V1, 2) V2&V3, 3) OPA, 4) EBA. Voxels inside these ROIs are assigned with an ideal value $\iota_i \in \{0, 0.33, 0.66, 1\}$. We fit a linear regression $\hat{\iota}_i = \beta \iota_i + \epsilon$, where slope β measures brain-model alignment, $b_0 = \epsilon$ and $b_1 = \beta + \epsilon$ measures the proportion of early and late layer not being selected.

We found that both qualitatively (Figure 5) and quantitatively (Table 2), layer-to-brain alignment is best in the CLIP model. Furthermore, the *hierarchy slope* increases as CLIP scaled up both model size and data (slope 0.32 for M, 0.50

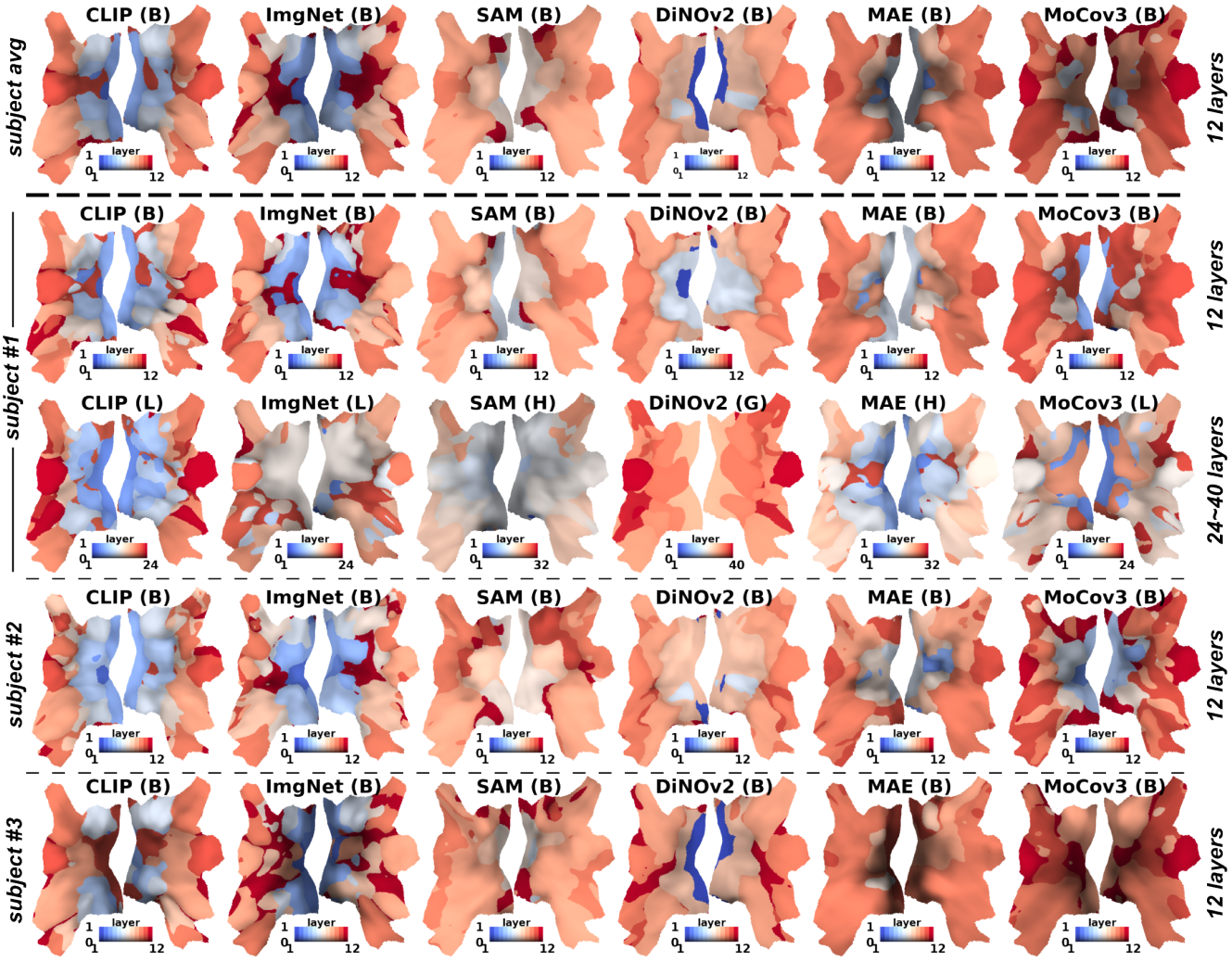

Figure 5. **Layer Selectors, Brain-Network Alignment**. All models are ViT architecture, number of layers is marked in the colorbar x-axis. Brightness is confidence measurement (defined in Section 3.2), and lower brightness means a *softer* selection of multiple layers. ***Top***: average of three subjects, base size 12-layer model. ***Middle***: subject #1, 12 layer small(S) and base(B) model, 24 layer large(L) model, 32 layer huge(H) model, 40 layer gigantic(G) model. ***Bottom***: subject #2 and #3, base size 12-layer model. ***Insights***: **1)** CLIP layers align best with the brain's hierarchical organization, **2)** ImageNet and SAM last layer align with mid-level in the brain, indicating their training objectives aimed at mid-level concept; **3)** DiNOv2: with a larger model, its hierarchy no longer align with the brain.

Model	**CLIP** [24]				**ImageNet** [11]		**SAM** [27]			**DiNOv2** [38]			**MAE** [21]			**MoCov3** [6]		
Size	L/14	B/16	B/32	B/32	L	B	H	L	B	G	L	B	H	L	B	L	B	S
Data	1B [15]	140M	14M	1.4M	IN-1K [11]		SA-1B [27]			LVD-142M [38]			IN-1K [11]			IN-1K [11]		
$R^2 \uparrow$	**0.132**	0.131	0.117	0.083	0.117	**0.121**	**0.120**	0.117	0.111	0.123	0.125	**0.128**	**0.132**	0.129	0.128	0.124	**0.127**	0.126
slope $\uparrow$	**0.53**	0.50	0.32	0.11	0.27	**0.39**	0.08	0.10	**0.15**	0.16	0.25	**0.41**	0.20	**0.37**	0.32	0.30	0.33	**0.40**
$b_0 \downarrow$	***0.35***	0.38	0.49	0.60	***0.41***	0.45	***0.58***	0.63	0.67	0.76	0.66	***0.50***	***0.46***	0.47	0.55	***0.40***	0.52	0.51
$b_1 \uparrow$	***0.88***	***0.88***	0.82	0.71	0.68	***0.83***	0.66	0.73	***0.82***	***0.92***	***0.92***	0.91	0.66	0.84	***0.87***	0.70	0.85	***0.91***

Table 2. **Layer Selectors, Brain-Network Alignment**. Brain-network alignment is measured by slope and intersection of linear fit (defined in Section 4.3). Larger **slope** means generally better alignment with the brain, smaller b_0 means better alignment of early layers, and larger b_1 means better alignment of late layers. R^2 is brain score. **Bold** marks the best within the same model. ***Insights***: **1)** CLIP's alignment to the brain improves with larger model capacity, **2)** for all others, bigger models decrease the brain-network hierarchy alignment.

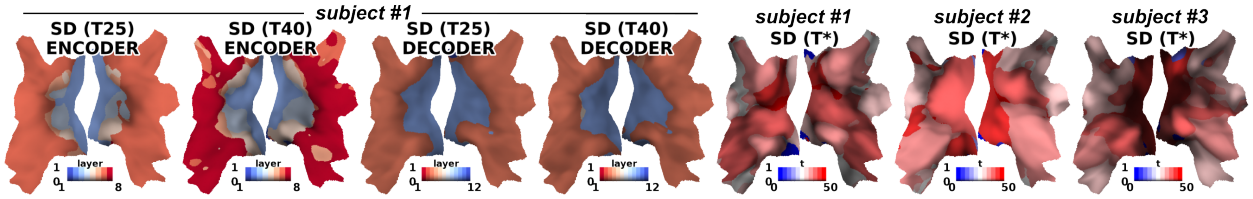

Figure 6. **Diffusion Models**. ***Left***: one time step (T=25 and T=40) layer selection, UNet encoder and decoder layers. ***Right***: fix layer (UNet decoder layer 6) time step selection. Color brightness is confidence measure (Section 3.2). ***Insights***: **1)** Diffusion models have less delineation in brain-network mapping using fixed time-step encoder/decoder layers, but more separation when using time steps, **2)** mid-late time steps align with higher level brain, late time step aligns with early brain region.

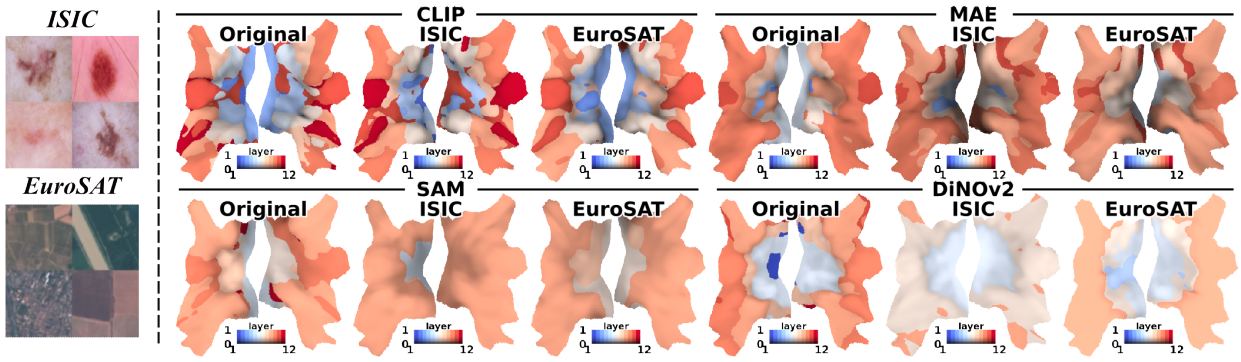

Figure 7. **fine-tuned to Small Datasets**. ***Left***: example images from ISIC and EuroSAT. ***Right***: layer selector (Colored as Section 3.2) before and after fine-tuning. The whole network is fine-tuned. ***Insights***: CLIP fine-tunes with less change in the existing computation.

for L, 0.52 for XL). CLIP M and S models were trained with the same model size but smaller data; the S model dropped *hierarchy slope* significantly (0.11). ImageNet, SAM, MoCov3, and DiNOv2 models show decreased *hierarchy slope* when scaling up: their late (or early for DiNOv2) layers were less selected for bigger models, indicating a decreasing hierarchical alignment with the brain.

4.4. Fine-tuned Model

The key findings are: **1)** CLIP maintains a hierarchical structure and uses less re-wiring for downstream tasks; **2)** DiNOv2 and SAM tend to re-wire their intermediate layers and lose their hierarchical structure rapidly when fine-tuned.

We fine-tuned on two small-scale downstream tasks, ISIC [9] skin cancer classification, and EuroSAT [23] satellite land-use classification. We used 50 training samples per class to train. The pre-trained model is fine-tuned across all layers with AdamW optimizer lr=3e-5, weight decay of 0.01, batch size of 4, for 3,000 steps. We verified that the fine-tuned models reached maximum validation performance without significant overfitting.

We apply brain-to-network mapping to visualize the fine-tuned networks. The first dataset ISIC skin cancer classification relies on low-level features. Figure 7 shows ImageNet/CLIP's last layer aligned with V1 after ISIC fine-tuning, potentially indicating the usage of top-down information for low-level vision tasks. The second dataset, EuroSAT, requires less fine-tuning on low-level features; V1 is still aligned to early layers for CLIP. After fine-tuning, qualitative results in Figure 7 showed CLIP and MAE maintained a strong hierarchical structure, while SAM and DiNOv2 largely lost their hierarchy; quantitative results in Table 3 showed brain score of CLIP and MAE dropped less compare to SAM and DiNOv2. Overall, CLIP and MAE adapt to dynamic tasks with less catastrophic forgetting and re-wiring of existing computation.

	Brain Score R^2 ↑		
Model / Fine-tune dataset	**Original**	**ISIC**	**EuroSAT**
CLIP	0.131	0.115	0.112
MAE	0.128	0.117	0.113
SAM	0.111	0.086	0.087
DiNOv2	0.128	0.085	0.082

Table 3. Brain score dropped after fine-tuning on small datasets.

4.5. Channels and Brain ROIs

The key findings are: **1)** we can cluster brain voxels using the co-occurrence of brain voxels with channels, and the clusters largely align well with known brain ROIs; **2)** we can compute brain ROI/cluster-specific responses on images to reveal the ROI functionality.

Recall our factorized multi-selector method compresses information across 4D network features into a channel-wise vector of $\boldsymbol{v}_i$ for each brain voxel i. Furthermore, channels

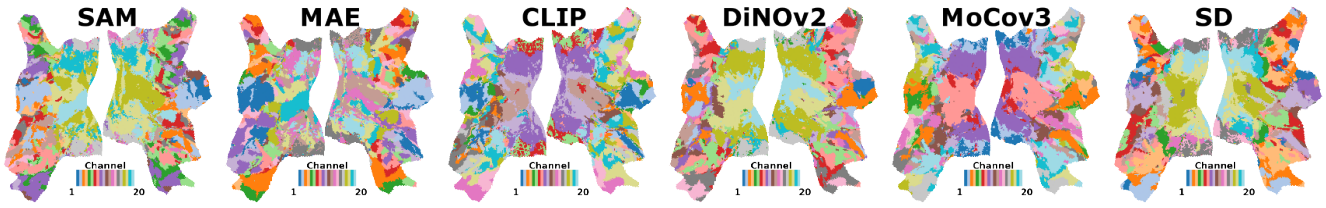

Figure 8. **Channel Clustering**. Brain voxels clustered by channel selection weight $\boldsymbol{w}_i$. ***Insights***: early visual brain uses less diverse channels but more diverse spatial locations (Figure 3), higher level brain is the opposite.

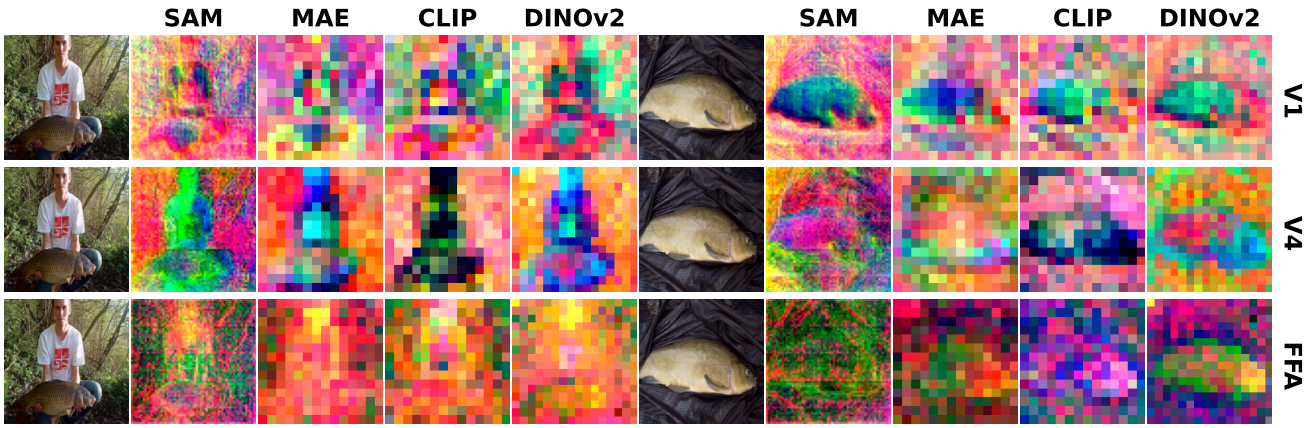

Figure 9. **Image features (selected channels).** Image RGB value corresponds to top-3 principal components (PCA) of n brain voxels' channel selection weights inside each brain ROI. The top-3 PCA channel selection weights are multiplied with channel-aligned image features and summed at every image pixel. **V1**: early visual brain, **V4**: mid-level visual brain, **FFA**: face-selective brain region.

across all the layers are aligned (Methods 3.1), resulting in a layer-agnostic channel representation. From that, a linear regression weight vector $\boldsymbol{w}_i$ acts as a *channel selector* to determine "*which feature channels best predict this brain voxel?*" We can view this channel feature selector, $\boldsymbol{w}(\boldsymbol{k})_i$, as co-occurrence between brain voxel i and channel elements k, which can be used to cluster the brain voxels: linking two voxels, i, j if they share similar channel selectors $\boldsymbol{w}_i$, $\boldsymbol{w}_j$. Figure 8 shows the result of clustering brain voxels into 20 clusters.

The higher-level brain utilized diverse channels across the brain areas; there is a consistent pattern that the face and body region use the same channel in CLIP, DiNOv2, MoCov3, and SD. The early visual brain used similar channels across the visual cortex; there is a consistent pattern that the left and right brain are symmetrical, as well as the ventral and parietal streams. SAM and MAE early visual brain-selected channels are non-symmetrical, indicating shift variant properties [57].

Furthermore, the selected channels reveal brain ROI's functionality. We visualized image feature response produced by the top-3 PCA components of channel weights within the selected ROIs (in Figure 9), which shows the brain ROIs encode low-level edge information in V1, mid-level semantic segmentation in V4, and face-selective features in FFA. Interestingly, DiNOv2 generalizes face across humans and fish [57, 64].

5. Discussion and Limitations

We have developed a visualization tool, *FactorTopy*, by training a robust brain encoding model. It allows us to see the internal working mechanism of any deep network. With this visualization and known functionality of brain ROIs, we can predict the network's downstream task performance, and diagnose their behavior when scaling up with a larger model or fine-tuning to a small dataset.

Limitations High-quality brain-encoding data of input images paired with brain fMRI responses is needed. NSD is the only such data publicly available. Over time, this situation might improve. Comparing brain-to-network alignments is less informative if networks' computation differs entirely from the brain. It is possible to achieve efficiency and generalization in a non-brain-like way; therefore, our tool is not universally applicable to all network designs.

Acknowledgement This work is supported by the funds provided by the National Science Foundation and by DoD OUSD (R&E) under Cooperative Agreement PHY-2229929 (The NSF AI Institute for Artificial and Natural Intelligence). Huzheng Yang and James Gee are supported in part by R01-HL133889, R01-EB031722, and RF1-MH124605.

References

[1] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, Jan. 2022. Number: 1 Publisher: Nature Publishing Group. 2, 3, 4

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, Dec. 2017. arXiv:1701.07875 [cs, stat]. 4

[3] Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, May 2019. Publisher: American Association for the Advancement of Science. 2

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021. arXiv:2104.14294 [cs]. 5

[5] Nadine Chang, John A. Pyles, Austin Marcus, Abhinav Gupta, Michael J. Tarr, and Elissa M. Aminoff. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, 6(1):49, May 2019. Number: 1 Publisher: Nature Publishing Group. 2

[6] Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers, Aug. 2021. arXiv:2104.02057 [cs]. 2, 4, 5, 6

[7] R. M. Cichy, K. Dwivedi, B. Lahner, A. Lascelles, P. Iamshchinina, M. Graumann, A. Andonian, N. A. R. Murty, K. Kay, G. Roig, and A. Oliva. The Algonauts Project 2021 Challenge: How the Human Brain Makes Sense of a World in Motion, Apr. 2021. arXiv:2104.13714 [cs, q-bio]. 2

[8] Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. The Algonauts Project: A Platform for Communication between the Sciences of Biological and Artificial Intelligence, May 2019. arXiv:1905.05675 [cs, q-bio]. 2

[9] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC), Mar. 2019. arXiv:1902.03368 [cs]. 7

[10] Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?, July 2023. Pages: 2022.03.28.485868 Section: New Results. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919. 4, 5, 6

[12] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, Feb. 2012. 2

[13] Serge O. Dumoulin and Brian A. Wandell. Population receptive field estimates in human visual cortex. *NeuroImage*, 39(2):647–660, Jan. 2008. 2

[14] Katrin Franke, Konstantin F. Willeke, Kayla Ponder, Mario Galdamez, Na Zhou, Taliah Muhammad, Saumil Patel, Emmanouil Froudarakis, Jacob Reimer, Fabian H. Sinz, and Andreas S. Tolias. State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930):128–134, Oct. 2022. Number: 7930 Publisher: Nature Publishing Group. 2

[15] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets, Oct. 2023. arXiv:2304.14108 [cs]. 4, 6

[16] Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, Dec. 2022. 2

[17] A. T. Gifford, B. Lahner, S. Saba-Sadiya, M. G. Vilas, A. Lascelles, A. Oliva, K. Kay, G. Roig, and R. M. Cichy. The Algonauts Project 2023 Challenge: How the Human Brain Makes Sense of Natural Scenes, Jan. 2023. arXiv:2301.03198 [cs, q-bio]. 2, 4

[18] Zijin Gu, Keith Wakefield Jamison, Meenakshi Khosla, Emily J. Allen, Yihan Wu, Ghislain St-Yves, Thomas Naselaris, Kendrick Kay, Mert R. Sabuncu, and Amy Kuceyeski. NeuroGen: Activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247:118812, Feb. 2022. 2

[19] Kamal Gupta, Gowthami Somepalli, Anubhav Anubhav, Vinoj Yasanga Jayasundara Magalle Hewa, Matthias Zwicker, and Abhinav Shrivastava. PatchGame: Learning to Signal Mid-level Patches in Referential Games. In *Advances in Neural Information Processing Systems*, volume 34, pages 26015–26027. Curran Associates, Inc., 2021. 2

[20] Matthew Gwilliam and Abhinav Shrivastava. Beyond Supervised vs. Unsupervised: Representative Benchmarking and Analysis of Image Representation Learning. pages 9642–9652, 2022. 2

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, Dec. 2021. arXiv:2111.06377 [cs]. 2, 4, 5, 6

[22] Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, Feb. 2023. Publisher: eLife Sciences Publications, Ltd. 2

[23] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A Novel Dataset and Deep Learn-

ing Benchmark for Land Use and Land Cover Classification, Feb. 2019. arXiv:1709.00029 [cs]. 7

[24] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, July 2021. 6

[25] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom Up Top Down Detection Transformers for Language Grounding in Images and Point Clouds. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 417–433, Cham, 2022. Springer Nature Switzerland. 2

[26] Nidhi Jain, Aria Wang, Margaret M. Henderson, Ruogu Lin, Jacob S. Prince, Michael J. Tarr, and Leila Wehbe. Selectivity for food in human ventral visual cortex. *Communications Biology*, 6(1):1–14, Feb. 2023. Number: 1 Publisher: Nature Publishing Group. 2

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything, Apr. 2023. arXiv:2304.02643 [cs]. 2, 4, 5, 6

[28] Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. Second Sight: Using brain-optimized encoding models to align image distributions with human brain activity, June 2023. arXiv:2306.00927 [cs, q-bio]. 2

[29] Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N. Apurva Ratan Murty, Kendrick Kay, Aude Oliva, and Radoslaw Cichy. BOLD Moments: modeling short visual events through a video fMRI dataset and metadata, Mar. 2023. Pages: 2023.03.12.530887 Section: New Results. 2

[30] Shamit Lal, Mihir Prabhudesai, Ishita Mediratta, Adam W. Harley, and Katerina Fragkiadaki. CoCoNets: Continuous Contrastive 3D Scene Representations. pages 12487–12496, 2021. 2

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, Feb. 2015. arXiv:1405.0312 [cs]. 4

[32] Andrew F. Luo, Margaret M. Henderson, Michael J. Tarr, and Leila Wehbe. BrainSCUBA: Fine-Grained Natural Language Captions of Visual Cortex Selectivity, Oct. 2023. arXiv:2310.04420 [cs, q-bio]. 2

[33] Andrew F. Luo, Leila Wehbe, Michael J. Tarr, and Margaret M. Henderson. Neural Selectivity for Real-World Object Size In Natural Images, Mar. 2023. Pages: 2023.03.17.533179 Section: New Results. 2

[34] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence, May 2023. arXiv:2305.14334 [cs]. 5

[35] Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay Jagadish, Eric Wang, Edgar Y. Walker, Santiago A. Cadena, Taliah Muhammad, Erick Cobos, Andreas S. Tolias, Alexander S. Ecker, and Fabian H. Sinz. Generalization in data-driven models of primary visual cortex. Jan. 2021. 2

[36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Aug. 2020. arXiv:2003.08934 [cs]. 2, 4

[37] Thomas P. O'Connell, Tyler Bonnen, Yoni Friedman, Ayush Tewari, Josh B. Tenenbaum, Vincent Sitzmann, and Nancy Kanwisher. Approaching human 3D shape perception with neurally mappable models, Sept. 2023. arXiv:2308.11300 [cs]. 2

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, Apr. 2023. arXiv:2304.07193 [cs]. 2, 4, 5, 6

[39] Jacob S. Prince, George A. Alvarez, and Talia Konkle. A contrastive coding account of category selectivity in the ventral visual stream, Aug. 2023. Pages: 2023.08.04.551888 Section: New Results. 2

[40] Jacob S. Prince, Ian Charest, Jan W. Kurzawski, John A. Pyles, Michael J. Tarr, and Kendrick N. Kay. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, 11:e77599, Nov. 2022. 4

[41] 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, Feb. 2021. arXiv:2103.00020 [cs]. 2, 4, 5

[42] N. Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J. DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature Communications*, 12(1):5540, Sept. 2021. Number: 1 Publisher: Nature Publishing Group. 2

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, Apr. 2022. arXiv:2112.10752 [cs]. 2, 4, 5

[44] Zvi N. Roth, Kendrick Kay, and Elisha P. Merriam. Natural scene sampling reveals reliable coarse-scale orientation tuning in human V1. *Nature Communications*, 13(1):6469, Oct. 2022. Number: 1 Publisher: Nature Publishing Group. 2

[45] Gabriel Sarch, Zhaoyuan Fang, Adam W. Harley, Paul Schydlo, Michael J. Tarr, Saurabh Gupta, and Katerina Fragkiadaki. TIDEE: Tidying Up Novel Rooms Using Visuo-Semantic Commonsense Priors. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 480–496, Cham, 2022. Springer Nature Switzerland. 2

[46] Gabriel Sarch, Hsiao-Yu Fish Tung, Aria Wang, Jacob

Prince, and Michael Tarr. 3D View Prediction Models of the Dorsal Visual Stream, Sept. 2023. arXiv:2309.01782 [cs, q-bio]. 2

[47] Gabriel H. Sarch, Michael J. Tarr, Katerina Fragkiadaki, and Leila Wehbe. Brain Dissection: fMRI-trained Networks Reveal Spatial Selectivity in the Processing of Natural Images, May 2023. Pages: 2023.05.29.542635 Section: New Results. 2

[48] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?, Sept. 2018. Pages: 407007 Section: New Results. 2

[49] Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 108(3):413–423, Nov. 2020. 2

[50] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models, Oct. 2022. arXiv:2210.08402 [cs]. 4

[51] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting Weakly Supervised Pre-Training of Visual Perception Models, Apr. 2022. arXiv:2201.08371 [cs]. 2

[52] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity, Nov. 2022. Pages: 2022.11.18.517004 Section: New Results. 2

[53] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent Correspondence from Image Diffusion, June 2023. arXiv:2306.03881 [cs]. 4

[54] JohnMark Taylor and Nikolaus Kriegeskorte. Extracting and visualizing hidden activations and computational graphs of PyTorch models with TorchLens. *Scientific Reports*, 13(1):14375, Sept. 2023. Number: 1 Publisher: Nature Publishing Group. 2

[55] Hsiao-Yu Fish Tung, Ricson Cheng, and Katerina Fragkiadaki. Learning Spatial Common Sense with Geometry-Aware Recurrent Networks, Apr. 2019. arXiv:1901.00003 [cs]. 2

[56] Polina Turishcheva, Paul G. Fahey, Laura Hansel, Rachel Froebe, Kayla Ponder, Michaela Vystrčilová, Konstantin F. Willeke, Mohammad Bashiri, Eric Wang, Zhiwei Ding, Andreas S. Tolias, Fabian H. Sinz, and Alexander S. Ecker. The Dynamic Sensorium competition for predicting large-scale mouse visual cortex activity from videos, May 2023. arXiv:2305.19654 [q-bio]. 2

[57] Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching Matters: Investigating the Role of Supervision in Vision Transformers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7486–7496, Vancouver, BC, Canada, June 2023. IEEE. 2, 4, 5, 8

[58] Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex, Sept. 2022. Pages: 2022.09.27.508760 Section: New Results. 2

[59] Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Pede, Max F. Burg, Christoph Blessing, Santiago A. Cadena, Zhiwei Ding, Konstantin-Klemens Lurz, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. The Sensorium competition on predicting large-scale mouse primary visual cortex activity, June 2022. arXiv:2206.08666 [cs, q-bio]. 2

[60] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models, Apr. 2023. arXiv:2303.04803 [cs]. 4

[61] Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, Mar. 2016. Number: 3 Publisher: Nature Publishing Group. 2

[62] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. Publisher: Proceedings of the National Academy of Sciences. 2

[63] Huzheng Yang, James Gee, and Jianbo Shi. Memory Encoding Model, Aug. 2023. arXiv:2308.01175 [cs]. 2

[64] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence, May 2023. arXiv:2305.15347 [cs]. 4, 8

[65] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer, Jan. 2022. arXiv:2111.07832 [cs]. 5

[66] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, Jan. 2021. Publisher: Proceedings of the National Academy of Sciences. 2