

Appendix

A1 Training Details

Consider a model type denoted by M . We train a pair of models, $\{M_1, M_2\}$ initialized with two different random seeds. We initialize the models using a uniform Xavier distribution [22]. This setup ensures that the two models are identical in architecture and achieve comparable task performance, allowing us to isolate the effects of stochastic variations in the SGD process (such as initialization differences and input order). By comparing the representations from these models, we can quantify the minimal set of transformations required to align them. Below, we outline the specific training parameters for each of the model types presented in Sec. 4.

Supervised Convolutional Neural Networks. All models are trained from scratch on CIFAR100 or ImageNet for 100 and 80 epochs respectively. We save model weights at every epoch and additionally store the best-performing weights based on test-set performance for each dataset.

Self-Supervised Networks. We train a pair of networks (ResNet50 backbone) using a Momentum Contrastive (MoCo) objective [23] initialized with 2 different random seeds on ImageNet for 50 epochs using a batch size of 256.

Vision Transformers. We analyze the ViT-B/16 variant of Vision Transformers (ViTs) [15] having patch size 16×16 trained on ImageNet and CIFAR100. We train multiple models using different random seeds for 25 epochs with batch size 32.

Language Models. We analyze models from the **Pythia suite** [6], a collection of autoregressive language models trained with varying architectures and random seeds. These models were predominantly trained on the **Pile dataset** [17]—a diverse and carefully curated corpus aggregating high-quality texts from sources such as academic publications, books, Wikipedia, and web-scraped data. This dataset provides a rich and heterogeneous distribution of language examples that supports robust learning of linguistic representations.

A2 Convergence with Network Depth Using Spearman’s Rank-Order Correlation

For analyses presented thus far in Sec. 4, we report alignment as a Pearson correlation. However, the use of such a metric could be susceptible to high-variance outlier dimensions. To address this possibility, we conduct an additional series of experiments, where we compute alignment scores using Spearman’s rank-correlation. Concretely, for an optimal transformation matrix M obtained after using a specific alignment metric (linear, Procrustes or permutation) to align a representational pair $\{X_i, X_j\}$, we now report Spearman’s rank coefficient between the aligned representations given by:

$$\text{Alignment} = \text{corr}(X_i, MX_j) = 1 - \frac{6 \sum (X_i - MX_j)}{n(n^2 - 1)}$$

where n is the number of stimuli.

As observed in Fig. A1, we note that the choice of correlation computation does not affect our conclusions.

A3 Basis Alignment in CNNs vs. Vision Transformers

In Sec. 4, an interesting phenomenon emerges—a privileged basis set persists in DCNNs, whereas in case of ViTs, there is no clear evidence of a privileged solution axis. While a definitive mechanistic account remains an open area of investigation to answer this question, recent work [29] offers compelling evidence that the emergence of basis alignment across CNNs and even between brains and CNNs—may be partly attributable to architectural choices, especially the presence of ReLU nonlinearities. To understand this constraint, we consider a representation of post-ReLU activations from a CNN, say x . The ReLU operation ensures that all activations $x \geq 0$, i.e.: non-negative. Now,

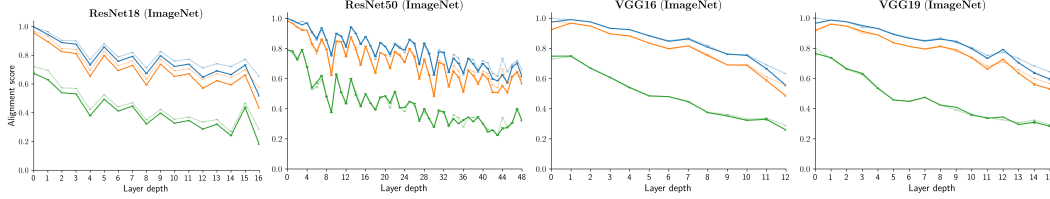


Figure A1: Representational Convergence Across a Network Hierarchy Using Spearman’s Rank Correlation. We plot the evolution of alignment scores computed between different (ImageNet-trained) network seeds of the same architecture. Lighter shades of the same color denote the Spearman’s ranked-correlation score, whereas darker shades indicates the Pearson correlation-based alignment score. We clearly observe that both correlation measures yield nearly identical results.

if we apply a random rotation, say Q to these activations, we obtain a rotated basis set $y = Qx$, where Q is a rotation matrix. For both x and y to be valid post-ReLU activations, they must remain non-negative after the transformation. In other words, we must strictly have $y \geq 0$. For this, the matrix Q must be a non-negative matrix. But this means that Q must be a permutation matrix, because every orthogonal matrix $Q \in O(N)$ with non-negative entries is necessarily a permutation matrix. Hence, it follows that Q can only permute (or shuffle) the activation units, rather than performing arbitrary rotations. Thus, the non-linearity induced by ReLU disrupts the rotational symmetry of the activation space, potentially explaining why different networks converge to similar bases. In contrast, Vision Transformers (ViTs) use GeLU nonlinearities in MLP layers. Moreover, the penultimate layer in ViTs often lacks *any* nonlinearity. These architectural choices retain greater rotational freedom in the feature space, which likely explains the lack of axis alignment across transformer runs, as also confirmed by our results in Section 4.5

A4 Additional Results Using CIFAR100

All analyses described in Sec. 4, we demonstrate evidence for representational convergence along the following directions—hierarchy effects, sensitivity to solution bases, hierarchical correspondence and training-time dynamics on ImageNet. We conduct an identical set of experiments on CIFAR100, and observe that our findings generalize across these datasets.

Network Hierarchy. We compare the representational convergence across a network hierarchy for different seeds of the same architecture using the CIFAR100 dataset in Fig. A2. This holds an

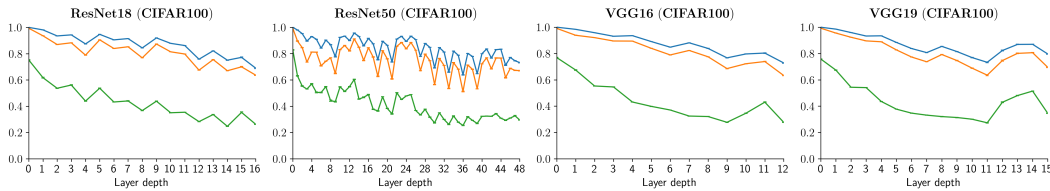


Figure A2: Representational Convergence Across a Network Hierarchy Using CIFAR100. Identical to ImageNet-trained networks, alignment follows the trend **Linear** > **Procrustes** > **Permutation**. Moreover, the **Linear** and **permutation** alignment scores track each other closely, again, identical to ImageNet-trained networks.

identical trend to those observed in ImageNet-trained networks—early layers show higher alignment, which tapers with network depth.

Sensitivity to Representational Axes. Identical to the procedure applied to ImageNet-trained networks in Sec. 4.1, we apply a random rotation matrix Q to the converged basis of a neural representation of CIFAR100-trained networks.

In Table A1, we note that alignment consistently decreases across all models after rotating the solution basis, identical to our observation on ImageNet networks.

Model	Native (Min / Max)	Rotated (Min / Max)	Difference (%) (Min / Max)
ResNet18	0.247 / 0.752	0.215 / 0.689	6.40% / 51.26%
ResNet50	0.254 / 0.828	0.242 / 0.828	-3.38% / 35.29%
VGG16	0.277 / 0.769	0.239 / 0.661	5.66% / 63.97%
VGG19	0.273 / 0.758	0.231 / 0.684	2.15% / 35.36%

Table A1: **Sensitivity of Permutation Scores to Representational Axes on CIFAR100.** For each CIFAR100-trained network we apply a random rotation to the network’s unit basis and recompute permutation alignment scores for all convolutional layers. Columns report the minimum and maximum alignment scores observed over layers in the native and rotated basis, and the final column gives the percentage change in alignment after rotation. Rotations reduce alignment, indicating that a privileged basis exists in trained networks independent of the training dataset.

Hierarchical Correspondence. We plot the heatmap of Procrustes and Soft-Matching alignment scores for all layer and network pairs using CIFAR100. For both alignment metrics, we observe

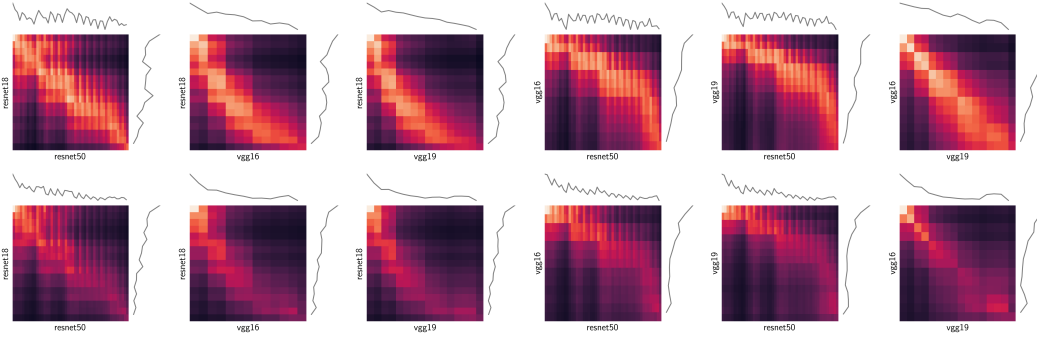


Figure A3: **Inter-Model Comparisons on CIFAR100.** We consider all pairs of vision models, and compute the alignment scores between every pair of layers using the orthogonal Procrustes (**Top**) and Soft-Matching (**Bottom**) metric trained on CIFAR100. Gray line plots denote the maximum alignment value for each network over rows and columns.

a hierarchical correspondence—layers at approximately similar depths in a network pair are more highly aligned than dissimilar depths.

Convergence over Training. We plot the Procrustes alignment scores between all network pairs trained on CIFAR100 from epochs 0 (untrained) through 10 in Fig. A4.

Identical to ImageNet networks in Sec. 4.2, we see that early convolutional layers have almost no alignment change, presumably due to the fact that early layers learn filters with approximately linear isometries. On the other hand, in later layers, we observe that the bulk of representational alignment happens in the first epoch itself, independent of network task performance.

A5 Out-Of-Distribution Datasets

All OOD datasets were directly taken from [20], which share the same 16 coarse labels as ImageNet. Concretely, this set consists of the following classes: Airplane, Bear, Bicycle, Bird, Boat, Bottle, Car, Cat, Chair, Clock, Dog, Elephant, Keyboard, Knife, Oven, Truck.

Each of the 17 stylized datasets are described below:

- **Color:** Half of the images are randomly converted to grayscale, and the rest kept in their original colormap.

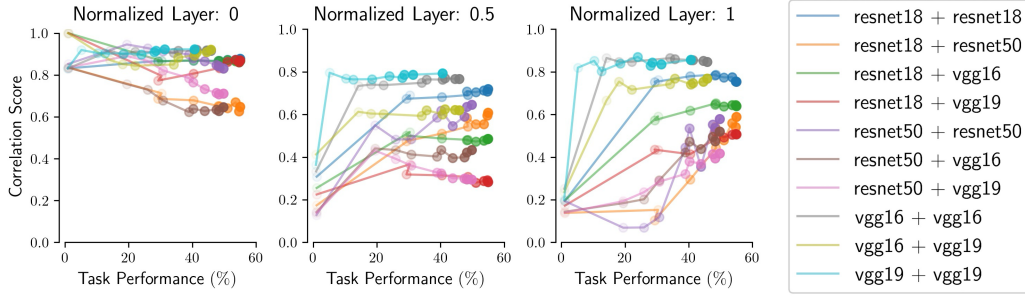


Figure A4: **Representational Alignment Through Training CIFAR100 Networks.** We plot the evolution of Procrustes alignment between network pairs during training on CIFAR100. Lighter shades indicate earlier epochs, progressively darkening with later epochs. The plots range from epoch 0 (untrained) to epoch 10, with task performance improving over time. Epoch progression can be inferred from the increasing task performance along the x -axis. This trend is identical to the convergence dynamics seen in ImageNet training—bulk of the alignment occurs within the first epoch itself, after which alignment saturates or even slightly reduces in some cases.

- **Stylized:** Textures from one class are transferred to the shapes of another, ensuring that object shapes remain preserved.
- **Sketch:** Cartoon-style sketches of objects representing each class.
- **Edges:** Generated from the original ImageNet dataset using the Canny edge detector to produce edge-based representations.
- **Silhouette:** Black objects on a white background generated from the original dataset.
- **Cue Conflict:** Images with textures that conflict with shape categories, generated using iterative style transfer [18], where **Texture** dataset images serve as the style and **Original** dataset images as the content.
- **Contrast:** Image variants modified to different contrast levels.
- **High-Pass / Low-Pass:** Images processed with Gaussian filters to emphasize either high-frequency or low-frequency components.
- **Phase-Scrambling:** Images with phase noise added to frequency components, introducing varying levels of distortion from 0° to 180° .
- **Power-Equalization:** The images were processed to normalize the power spectra across the dataset by adjusting all amplitude spectra to match the mean value.
- **False-Color:** The colors of the images were inverted to their opponent colors while maintaining constant luminance, using the DKL color space.
- **Rotation:** Rotated images (0° , 90° , 180° , or 270°) to test rotational invariance.
- **Eidolon I, II, III:** The images were distorted using the Eidolon toolbox, with variations in the coherence and reach parameters to manipulate both local and global image structures for each intensity level.
- **Uniform Noise:** White uniform noise was added to the images in a varying range to assess robustness, with pixel values exceeding the bounds clipped to the range $[0, 255]$.

A6 Additional Results on Convergence Across Distribution Shifts

We also computed Procrustes alignment for the remainder of vision networks at the first convolutional and penultimate layer to assess whether a similar phenomenon holds as described in Sec. 4.3. Indeed, in Fig. A5 we observe a similar trend that was observed earlier, i.e.: alignment mirrors task performance at higher network depths.

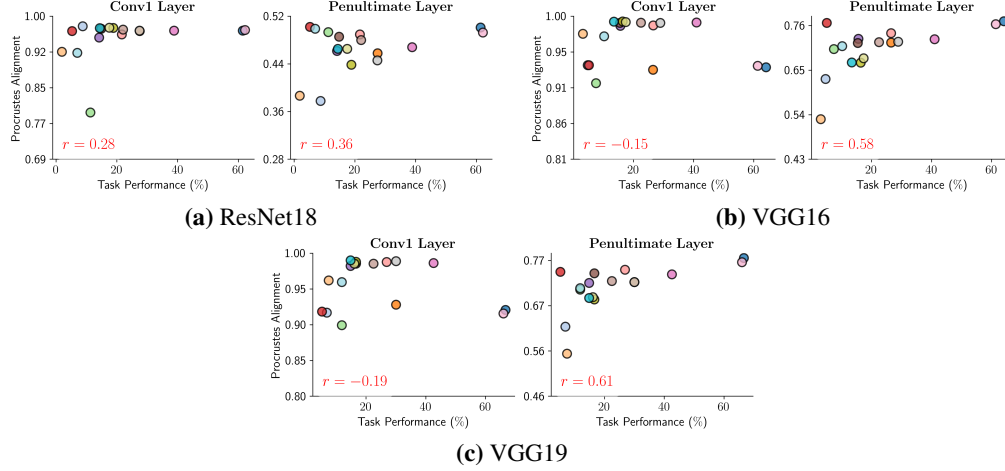


Figure A5: **Procrustes Alignment vs. Task Performance** We compute the Procrustes alignment of different network architectures on each of the 17 datasets for the first convolutional layer (**Left**) and the penultimate (**Right**) layer from (a) - (c).

A7 Representational Alignment Over Training

In Section 4.2 we compared networks trained for identical epochs and found that representational alignment plateaued within the first epoch. This rapid convergence, however, could still reflect networks following similar developmental trajectories driven by task optimization—essentially reaching high alignment early because they traverse a universal learning path toward the task solution.

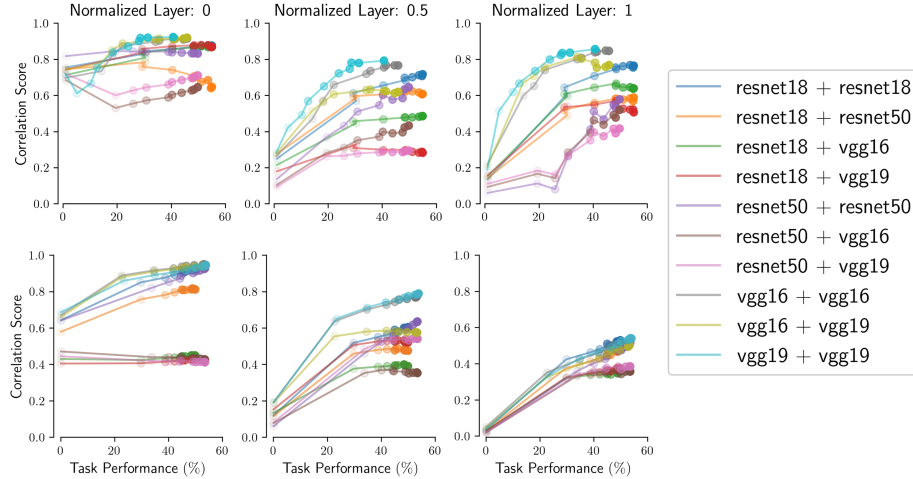


Figure A6: **Evolution of Representational Alignment to a Fully Trained Network.** Procrustes alignment between each training checkpoint and the fully trained reference model, shown for CIFAR-100 (**Top**) and ImageNet (**Bottom**). Each marker is one epoch (0 = untrained, 10 = ten epochs), with color lightening for early epochs and darkening as training progresses. Alignment climbs sharply within the first epoch and then levels off, while the earliest convolutional layers exhibit only minimal change—highlighting that most convergence occurs long before peak task performance is reached.

To test whether task-optimization explains this phenomenon, we compared fully trained networks with networks at various intermediate training stages. Remarkably, high representational alignment still emerged predominantly within the first epoch, well before networks achieved optimal task

performance (Fig. A6). The earliest convolutional layers showed minimal change throughout training, consistent with learning approximately linear transformations for basic visual feature extraction. Altogether, these results imply that representational convergence is driven by early optimization dynamics, not by attaining the final task solution.

A8 Comparisons to Brain Data

In the following section, we apply our comparative analysis framework on brain data (Sec. I). We analyze fMRI responses from four subjects (IDs 1, 2, 5 and 7) using data from the Natural Scenes Dataset (NSD) [1]. In this dataset, each subject viewed 37,000 *naturalistic* images, with 1000 images shared among all participants. For our analysis, we use these 1000 shared images to find how representational alignment between different subjects brains changes across the network hierarchy and to better understand the minimal sets of transformations needed to align two brains. We use the Soft-Matching score instead of the permutation alignment score since the number of recorded voxels is different across all subjects.

We align responses from five key brain regions along the visual pathway: V1, V2, V3, V4, and the high-level ventral stream, arranged in approximate order of increasing visual processing depth. Regions V1–V4 are defined using the population receptive field (pRF) localizer scan session from the NSD, and the high-level ventral visual stream region is delineated according to the NSD streams atlas. All alignment values are normalized by the mean noise ceiling for each brain region, with noise ceilings computed following the standard procedure described in [1], based on the variability in voxel responses across three repeat measurements per stimulus.

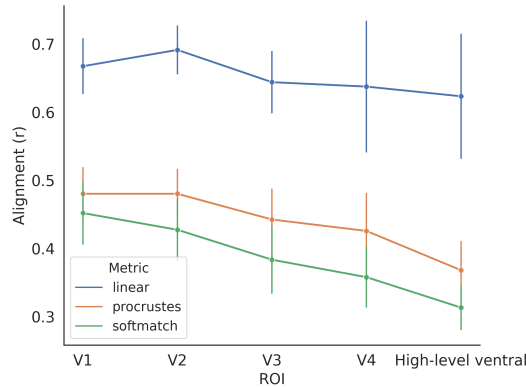


Figure A7: **Convergence Across the Visual Cortex.** Evolution of alignment scores computed between different NSD participants across the visual cortex hierarchy. Consistent with Fig. I, alignment decreases along the *depth* of the visual cortex. Notably, **Soft-Matching** achieves comparable alignment scores to **Procrustes**, suggesting a strong, region-specific voxel correspondence across subjects. Error bars denote standard deviation across all ($n = 6$) participant pairs.

We observe that consistent with network results, inter-subject alignment decreases with visual processing depth across all alignment metrics (Fig. A7). However, unlike the network results, the soft-match scores closely approximate Procrustes scores in these brain data, suggesting that voxel responses are already highly axis-aligned across subjects and thus leave little room for rotations to further improve alignment. Notably, we also observe a substantial gap between Procrustes and linear alignment in the brain data, in contrast to ANNs where Procrustes closely approximates linear alignment. This discrepancy implies that inter-subject variability in human brains requires more flexible transformations (*e.g.*, scaling or shearing) to achieve high alignment. Such variability could stem from individual differences in anatomical and functional organization, or from imperfect cross-subject ROI definitions.

A9 Choosing Random Pixels in the Convolutional Map

Throughout the manuscript, we use the central pixel from each convolutional feature map as a representative sample for alignment analyses. However, this begs a simple question—does spatial choice bias our results? To test this, we repeat the analyses using a random activation pixel for each model–seed pair in ImageNet-trained networks (Fig. A8). Through this experiment, we see that in

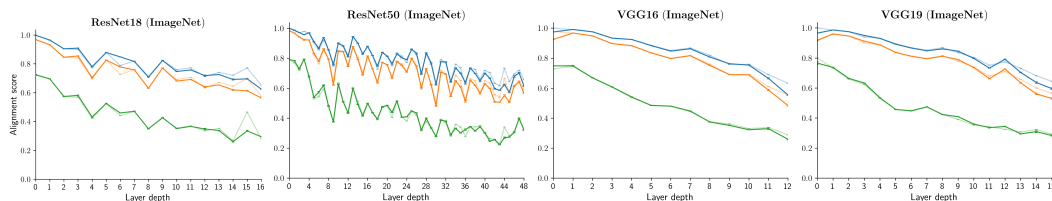


Figure A8: **Representational Convergence Across a Network Hierarchy Using a Random Pixel.** We plot the representational convergence across (ImageNet-trained) network hierarchies using both—the central pixel (darker shade) and a random pixel (lighter shade) for 3 alignment metrics—[Linear predictivity](#), [Procrustes](#), [Permutation](#). Across all these metrics, we observe that the spatial choice of the sample pixel leaves the alignment effectively unchanged.

fact choosing an arbitrary spatial location results in alignment trends across the network hierarchy remaining effectively unchanged, confirming that the choice of spatial location does not qualitatively affect our conclusions. Although using the full spatial map would be ideal, it is computationally prohibitive—scaling polynomially with dataset size—making the single-pixel approach an efficient and reliable proxy.

References

- [1] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.
- [2] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019.
- [3] A. Atanasov, B. Bordelon, and C. Pehlevan. Neural networks as kernel learners: The silent alignment effect. *arXiv preprint arXiv:2111.00034*, 2021.
- [4] Y. Bansal, P. Nakkiran, and B. Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.
- [5] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.
- [6] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O’Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.
- [7] Y. Bo, A. Soni, S. Srivastava, and M. Khosla. Evaluating representational similarity measures from the lens of functional correspondence. *arXiv preprint arXiv:2411.14633*, 2024.
- [8] L. Braun, E. Grant, and A. M. Saxe. Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. In *Forty-second International Conference on Machine Learning*, 2025.
- [9] R. Cao and D. Yamins. Explanatory models in neuroscience: Part 2—constraint-based intelligibility. *arXiv preprint arXiv:2104.01489*, 2021.

- [10] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.
- [11] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):27755, 2016.
- [12] C. Conwell, J. S. Prince, K. N. Kay, G. A. Alvarez, and T. Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1):9383, 2024.
- [13] J. A. De Loera and E. D. Kim. Combinatorics and geometry of transportation polytopes: An update. *Discrete geometry and algebraic combinatorics*, 625:37–76, 2013.
- [14] C. C. Dominé, N. Anguita, A. M. Proca, L. Braun, D. Kunin, P. A. Mediano, and A. M. Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. *arXiv preprint arXiv:2409.14623*, 2024.
- [15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [16] J. Frankle, D. J. Schwab, and A. S. Morcos. The early phase of neural network training. *arXiv preprint arXiv:2002.10365*, 2020.
- [17] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- [18] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.
- [19] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.
- [20] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.
- [21] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.
- [22] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.
- [23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [25] M. Huh, B. Cheung, T. Wang, and P. Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.
- [26] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.
- [27] S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

- [28] M. Khosla and A. H. Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pages 326–341. PMLR, 2024.
- [29] M. Khosla, A. H. Williams, J. McDermott, and N. Kanwisher. Privileged representational axes in biological and artificial neural networks. *bioRxiv*, pages 2024–06, 2024.
- [30] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.
- [31] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [32] J. Kubilius, M. Schrimpf, K. Kar, R. Rajalingham, H. Hong, N. Majaj, E. Issa, P. Bashivan, J. Prescott-Roy, K. Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.
- [33] Y. Li, J. Yosinski, J. Clune, H. Lipson, and J. Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.
- [34] J. Mehrer, C. J. Spoerer, N. Kriegeskorte, and T. C. Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, 2020.
- [35] A. Morcos, M. Raghu, and S. Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.
- [36] L. Moschella, V. Maiorca, M. Fumero, A. Norelli, F. Locatello, and E. Rodolà. Relative representations enable zero-shot latent space communication (2023). *arXiv preprint arXiv:2209.15430*, 2023.
- [37] A. Prasad, U. Manor, and T. Pereira. Exploring the role of image domain in self-supervised dnn models of rodent brains. In *SVRHM 2022 Workshop@ NeurIPS*, 2022.
- [38] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.
- [39] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.
- [40] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.
- [41] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.
- [42] M. Schrimpf, J. Kubilius, M. J. Lee, N. A. R. Murty, R. Ajemian, and J. J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.
- [43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [44] I. Sucholutsky, L. Muttenthaler, A. Weller, A. Peng, A. Bobu, B. Kim, B. C. Love, E. Grant, I. Groen, J. Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.
- [45] A. H. Williams, E. Kunz, S. Kornblith, and S. Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021.
- [46] D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

- [47] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.
- [48] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.
- [49] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.
- [50] C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, and D. L. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.