

## Appendix

### A1 Network Training

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 [23]. 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. 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.

### A2 Out-Of-Distribution Datasets

All OOD datasets were directly taken from [21], 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.
- **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 [19], 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  $0^\circ$  to  $180^\circ$ .
- **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 ( $0^\circ$ ,  $90^\circ$ ,  $180^\circ$ , or  $270^\circ$ ) 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  $[0, 255]$ .

### A3 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. A1 we observe a similar trend that was observed earlier, i.e., alignment mirrors task performance at higher network depths.

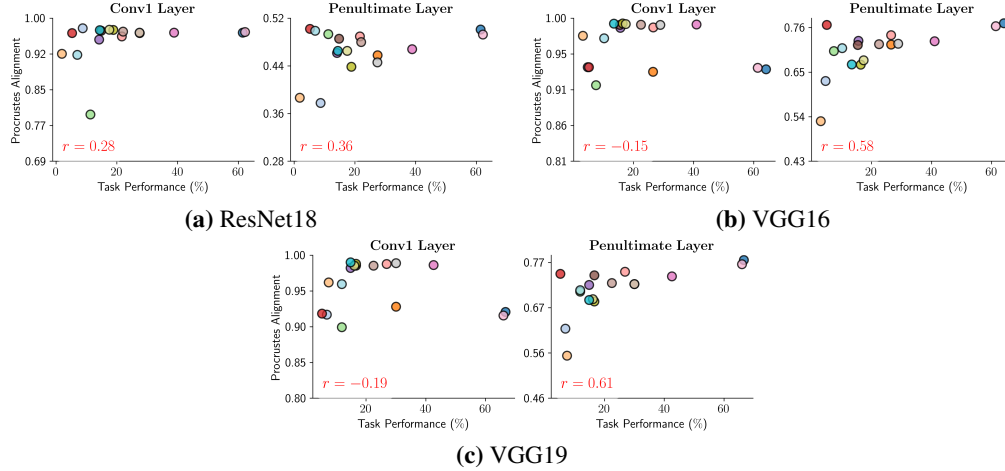


Figure A1: **Procrustes alignment vs. task performance** We compute the Procrustes alignment of different network architectures on each of the 17 datasets for the first convolutional layer (**Left**) and the penultimate (**Right**) layer from (a) - (c).

#### 444 A4 Representational Alignment Over Training

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

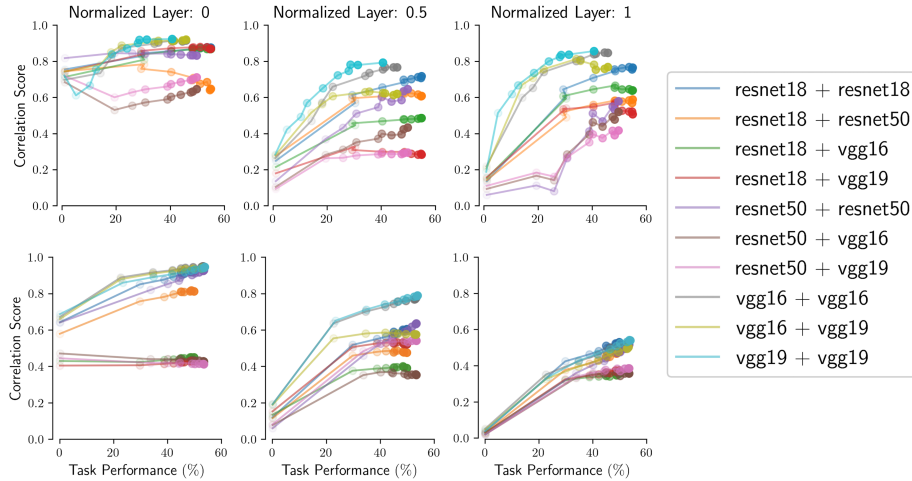


Figure A2: **Evolution of Representational Alignment to a Fully Trained Network.** Procrustes alignment between each training checkpoint and the fully trained reference model, shown for CIFAR-100 (top) and ImageNet (bottom). Each marker is one epoch (0 = untrained, 10 = ten epochs), with color lightening 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.

449

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

performance (Fig. A2). 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.

## A5 Representational Alignment for Self-Supervised Networks

So far, we have demonstrated our findings using a fully supervised learning objective. To assess whether these findings generalize to alternative training paradigms, we next investigate the effect of a self-supervised learning approach. In particular, we validate our results on models trained with an unsupervised framework using a contrastive learning objective—Momentum Contrast (MoCo) [24]. We train a pair of networks (ResNet50 backbone) initialized with 2 different random seeds on ImageNet for 50 epochs.

**Alignment Across Network Hierarchy.** Much like fully supervised networks, our results indicate that representational convergence is most pronounced in early layers and decreases with network depth (Fig. A3-A). We also note an analogous trend of the linear and Procrustes alignment scores being comparable, suggesting that representational variability across models that can be explained well using rotations and reflections, rather than complex affine transformations.

**Convergence Across Distribution Shifts.** When comparing representational alignment between self-supervised networks across distribution shifts, we observe that OOD and in-distribution alignment values are closely matched in early layers, indicating shared representational structure robust to distribution shifts. In contrast, deeper layers show increasing divergence (Fig. A3-B) for OOD inputs. This echoes our findings from supervised networks—even under self-supervised learning, networks initially develop general-purpose filters that are fairly aligned across all stimulus distributions but progressively evolve task-specific representations sensitive to distribution shifts.

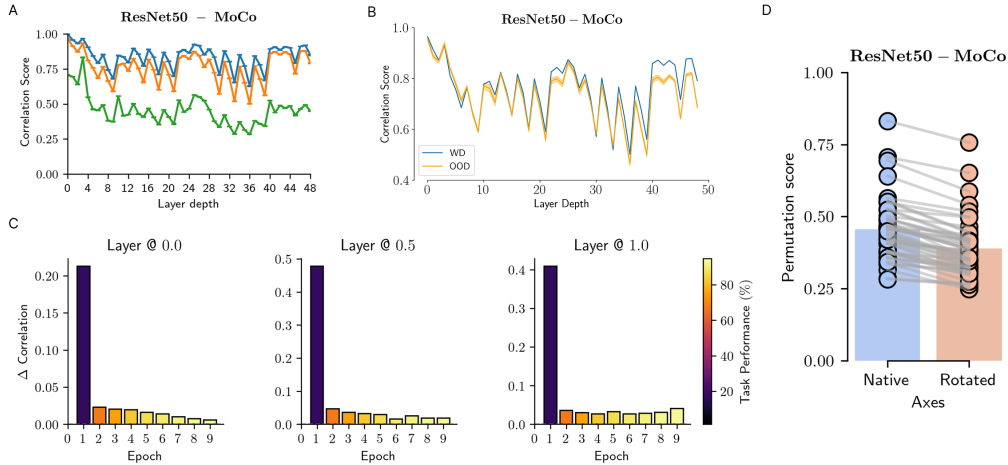


Figure A3: **Representational Alignment with Self-Supervised Networks.** (A) Representational alignment across layers of a pair of MoCo-trained ResNet50 models on ImageNet. Error bars indicate standard deviation from 5-fold cross-validation. (B) We plot the Procrustes alignment between MoCo models evaluated on in-distribution (ImageNet) and out-of-distribution (Stylized ImageNet [21]) stimuli. Error bars show standard error across all ( $n = 17$ ) OOD datasets. (C) Change in Procrustes alignment score with every training epoch. Colors in the bar-plot indicate the top-1 accuracies during training. (D) Permutation alignment in the native and randomly rotated basis. Each dot corresponds to a convolutional ResNet50 layer. Bars indicate mean alignment. Rotation reduces mean alignment by 14.87%.

**Alignment Over Training.** In Fig. A3-C, we show how the correlation score computed using the Procrustes alignment metric for a pair of networks evolves over training. Consistent with our

findings in the supervised regime (Sec. 4.2), the bulk of representational alignment occurs in the first epoch—well before task performance has peaked.

**Sensitivity To Permutation Alignment.** We apply a random rotation to the converged basis of MoCo-trained models. Consistent with our earlier findings, this perturbation leads to a significant drop in permutation alignment (Fig. A3-D). These results suggest that regardless of the learning paradigm, networks converge to a privileged basis.

## A6 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 1,000 images shared among all participants. For our analysis, we use these 1,000 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.

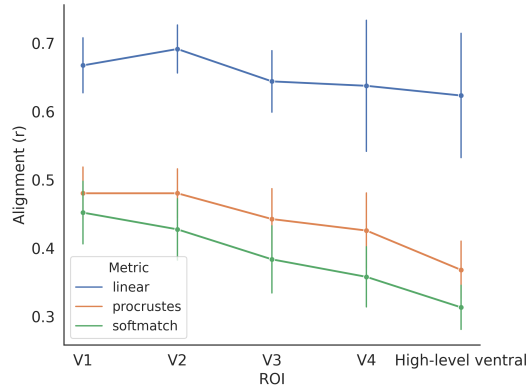


Figure A4: **Convergence Across the Visual Cortex.** Evolution of alignment scores computed between different NSD participants across the visual cortex hierarchy. Consistent with Fig. I, alignment decreases along the *depth* of the visual cortex. Notably, **Soft-Matching** achieves comparable alignment scores to **Procrustes**, suggesting a strong, region-specific voxel correspondence across subjects. Error bars denote standard deviation across all ( $n = 6$ ) participant pairs.

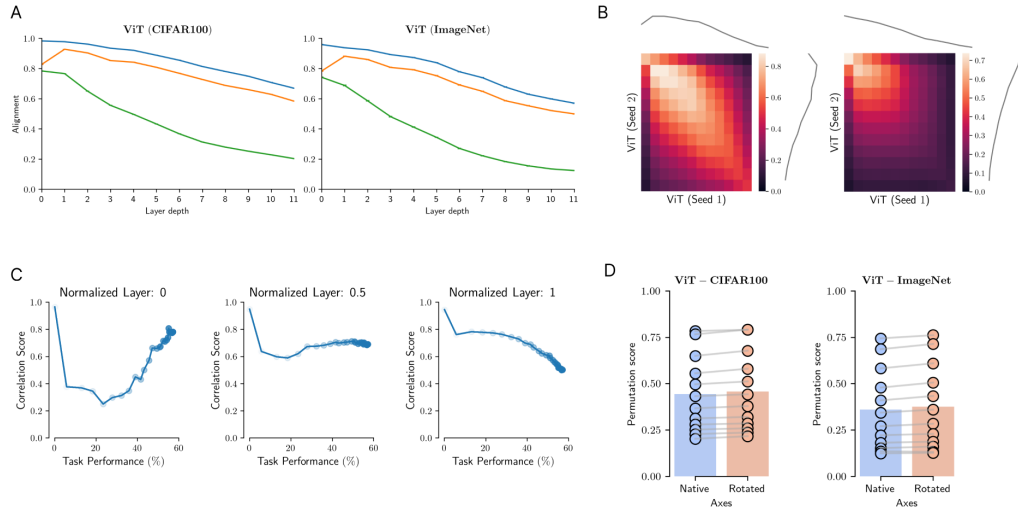
We observe that consistent with network results, inter-subject alignment decreases with visual processing depth across all alignment metrics (Fig. A4). However, unlike the network results, the soft-match scores closely approximate Procrustes scores in these brain data, suggesting that voxel responses 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.

## A7 Representational Convergence in Vision Transformers

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

**Computing Alignment.** For ViTs models, we analyze the [CLS] (class) token representations at each layer. This is because the [CLS] token aggregates the “information content” of all patch embeddings via the self-attention operator, making it a useful proxy to study representational geometry. In addition, computing representational similarity for [CLS] vectors avoids having to deal with pooling patch similarities into a single representative score.

**Results.** We conduct analyses akin to DCNNs in Sec. 4. This yields several key insights on the following fronts:



**Figure A5: Representational Convergence in Vision Transformers.** (A) We plot the evolution of alignment scores of three metrics (Linear, Procrustes, Permutation) computed between different seeds of the same ViT, which was trained on CIFAR100 (left) and ImageNet (right). Error bars denote the standard deviation across 5-fold cross validation. (B) We plot the inter-model orthogonal Procrustes (left) and permutation (right) scores for all layer pairs. Gray line plots denote the maximum alignment value over rows (right line) and columns (top line). (C) We visualize the evolution of the orthogonal Procrustes score at different checkpoints, ranging from epochs 0 (untrained) to 25. Darker colors correspond to epoch progression. (D) We rotate the converged basis of ViTs by a random rotation matrix and recompute the permutation scores for models trained on CIFAR100 (left) and ImageNet (right). Each dot represents a layer in the ViT. The permutation alignment remains approximately constant in both cases.

- Convergence Over Network Depth:** Identical to DCNNs (Sec. 4.1), we observe a consistent downward trend of alignment scores with network depth across all metrics in Fig. A5-A. In addition, the Procrustes and linear alignments follow each other closely, again, suggesting that rotations and reflections can explain much of the variance in learned representations of ViTs. Further, the layer-wise alignment computed using the Procrustes metric reveals a pattern similar to CNNs (Fig. A5-B, left).
- Absence of privileged axes in ViTs:** Unlike DCNNs, independent runs of the same ViT architecture do not converge to similar bases, and permutation scores between native and rotated axes remain statistically indistinguishable across all layers. This indicates that ViTs do not develop consistent, shared axes of representation across training seeds (Fig. A5-D).
- Early plateau of alignment in ViTs:** Consistent with DCNNs, alignment across different ViT seeds shows no systematic increase after epoch 1 for any layer except the input embed-

ding, implying that late-stage training does not drive convergence (Fig. A5-C). The [CLS] token’s apparent alignment spike at initialization is likely an artifact of its uniform positional encoding and weight initialization rather than true inter-seed convergence.

## A8 Representational Convergence in Language Models

**Model and Dataset Details.** 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** [18]—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.

Specifically, our analyses includes:

- **Same-Architecture Comparisons:** We compare multiple instances of the **Pythia-160m** model, which share an identical architecture but differ in the initialization of random seeds. This allows us to assess the variability in representational spaces resulting solely from stochastic training factors.
- **Cross-Architecture Comparisons:** In addition to seed variation, we compare models with different architectural configurations, namely the **Pythia-70m** and the **Pythia-160m** models. The primary architectural difference lies in network depth (i.e., number of layers), offering insights into how differences in model capacity and depth impact the learned representations.

All alignment metrics are computed using the representations of all unique sentences from the Semantic Textual Similarity Benchmark (STSB) dataset [10].

**Methodology.** For our representational similarity analysis, we perform two sets of computations:

**1. Final Checkpoint Analysis:** For each model at its final checkpoint, we compute the similarity between every pair of layers using the three metrics used for vision models, namely the Permutation, Procrustes and Linear Alignment measures.

**2. Intermediate Checkpoint Analysis:** Across the training run, we analyze 154 intermediate checkpoints sampled during the training process of the Pythia suite. At each checkpoint, for each model, we extract layer representations at normalized indices corresponding to the beginning (0), middle (0.5), and end (1) of the network. The similarity between these layers is then computed using the **Procrustes** measure.

**Results.** Comparing representations across different language models yields several key insights:

- **Metric-Dependent Alignment:** As shown in Figure A6-A, the evolution of alignment scores across the network hierarchy reveals that linear alignment consistently yields higher similarity scores compared to the Procrustes measure, which in turn scores higher than the permutation metric. However, as observed with vision networks, the Procrustes alignment approaches the linear alignment score implying that simple rotational and reflectional transformations account for much of the variability in representations across different language models. Further, unlike the vision case, high alignment is maintained across the network hierarchy and does not diminish with network depth.
- **Hierarchical correspondence:** Layer-wise alignment comparisons reveal a pattern analogous to the one observed in vision models: early layers correspond most closely with early layers, mid-level layers with mid-level ones, and late layers with their counterparts. This systematic correspondence is evident both—for models sharing the same architecture (Figure A6-B) and for models with different architectures (Figure A6-C).
- **Rapid representational convergence** An examination of intermediate checkpoints shows that representational alignment emerges early for all three key layers (normalized indices 0, 0.5, and 1). Alignment scores generally peak at around 8 – 10 billion tokens (Figure A6-D), after which they stabilize or decline slightly—even as next-word prediction performance



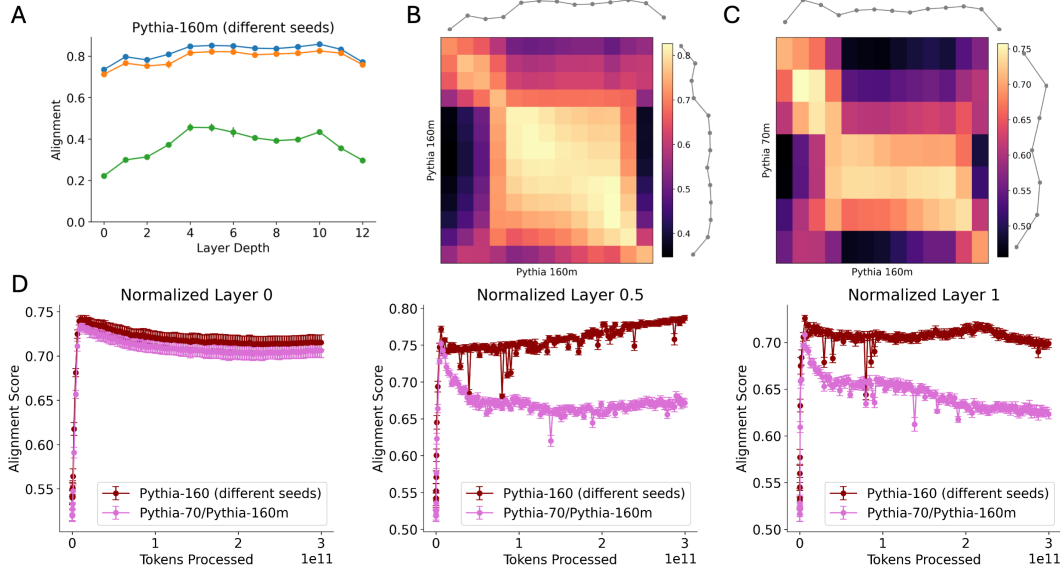


Figure A6: **Representational convergence in Language Models.** (A) We plot the evolution of alignment scores (computed between different seeds of the same network architecture) across the network hierarchy for language models (Pythia-160m) trained with different seeds on the same Pile dataset. The alignment values are computed using all sentences from the Semantic Textual Similarity Benchmark [10]. Alignment consistently follows the order: **Linear** > **Procrustes** > **Permutation**, reflecting the progressively stricter nature of the metrics. Notably, **Procrustes** transformations align representations nearly as well as **Linear** transformations, mimicking the trend observed for vision models. (B) Inter-Model Orthogonal Procrustes. We consider pairs of Pythia-160m models trained on different seeds, and for each pair, compute the alignment scores between every pair of layers using the orthogonal Procrustes metric. (C) Same as B but computed between networks with different architectures (specifically layer depths). (D) We visualize the evolution of Procrustes alignment between network pairs during across different model checkpoints. Alignment peaks around rapidly 8-10B tokens before saturating or declining even though the next-word prediction performance continues to improve.

continues to improve up to 300B tokens as reported in the Pythia paper [6]. This mirrors the evolution observed in vision models.

## 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] R. Burkard, M. Dell’Amico, and S. Martello. *Assignment problems: revised reprint*. SIAM, 2012.

[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] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[15] 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.

[16] 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.

[17] J. Frankle, D. J. Schwab, and A. S. Morcos. The early phase of neural network training. *arXiv preprint arXiv:2002.10365*, 2020.

[18] 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.

[19] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[20] 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.

[21] 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.

[22] 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.

[23] 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.



- [24] 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.
- [25] 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.
- [26] M. Huh, B. Cheung, T. Wang, and P. Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.
- [27] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.
- [28] 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.
- [29] 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.
- [30] M. Khosla, A. H. Williams, J. McDermott, and N. Kanwisher. Privileged representational axes in biological and artificial neural networks. *bioRxiv*, pages 2024–06, 2024.
- [31] 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.
- [32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [33] 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.
- [34] 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.
- [35] J. Mehrer, C. J. Spoerer, N. Kriegeskorte, and T. C. Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, 2020.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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.

- 693 [43] M. Schrimpf, J. Kubilius, M. J. Lee, N. A. R. Murty, R. Ajemian, and J. J. DiCarlo. Integra-  
694 tive benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*,  
695 108(3):413–423, 2020.
- 696 [44] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image  
697 recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- 698 [45] I. Sucholutsky, L. Muttenthaler, A. Weller, A. Peng, A. Bobu, B. Kim, B. C. Love, E. Grant,  
699 I. Groen, J. Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint*  
700 *arXiv:2310.13018*, 2023.
- 701 [46] A. H. Williams, E. Kunz, S. Kornblith, and S. Linderman. Generalized shape metrics on neural  
702 representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021.
- 703 [47] D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory  
704 cortex. *Nature neuroscience*, 19(3):356–365, 2016.
- 705 [48] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-  
706 optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of*  
707 *the national academy of sciences*, 111(23):8619–8624, 2014.
- 708 [49] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural  
709 networks? *Advances in neural information processing systems*, 27, 2014.
- 710 [50] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer*  
711 *Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014,*  
712 *Proceedings, Part I 13*, pages 818–833. Springer, 2014.
- 713 [51] C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, and D. L. Yamins.  
714 Unsupervised neural network models of the ventral visual stream. *Proceedings of the National*  
715 *Academy of Sciences*, 118(3):e2014196118, 2021.