3D View Prediction Models of the Dorsal Visual Stream

Gabriel H. Sarch (gsarch@andrew.cmu.edu)
Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA.

Hsiao-Yu Fish Tung (hytung@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA.

Aria Y. Wang (ariawang@cmu.edu)
Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA.

Jacob S. Prince (jacob.samuel.prince@gmail.com)
Department of Psychology, Harvard University, Cambridge, MA, USA.

Michael J. Tarr (michaeltarr@cmu.edu)
Department of Psychology & Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA.

Abstract

Deep neural network representations align well with brain activity in the ventral visual stream. However, the primate visual system has a distinct dorsal processing stream with different functional properties. To test if a model trained to perceive 3D scene geometry aligns better with neural responses in dorsal visual areas, we trained a self-supervised geometry-aware recurrent neural network (GRNN) to predict novel camera views using a 3D feature memory. We compared GRNN to self-supervised baseline models that have been shown to align well with ventral regions using the large-scale fMRI Natural Scenes Dataset (NSD). We found that while the baseline models accounted better for ventral brain regions, GRNN accounted for a greater proportion of variance in dorsal brain regions. Our findings demonstrate the potential for using task-relevant models to probe representational differences across visual streams.

Keywords: Visual Streams. DNN. Self-Supervision. fMRI.

The visual cortex has been traditionally organized into two processing streams (Ungerleider, 1982), the ventral and dorsal[1] (parietal) pathways, with a third lateral pathway being proposed recently (Weiner & Grill-Spector, 2013; Wurm & Caramazza, 2022; Pitcher & Ungerleider, 2021). Deep neural networks (DNNs) trained for object recognition have been found to be highly predictive of the ventral visual stream processing (Yamins et al., 2014). However, it remains unclear whether DNNs for recognition are well suited for predicting non-ventral visual processing, in the lateral or parietal visual streams.

DNNs optimized for egomotion estimation or action recognition may better predict neural responses in the parietal and lateral visual streams (Mineault et al., 2021; Güçlü & van Gerven, 2017). However, DNNs trained for action recognition do not appear to differentiate themselves from DNNs trained for object recognition in terms of predicting activity across the visual streams (Finzi et al., 2022). Recent experimental work has demonstrated evidence that the parietal stream plays a major role in global shape perception during object recognition, while the ventral stream may be more involved in local shape and texture encoding (Ayzenberg & Behrmann, 2022). Additionally, a well-established function of the parietal pathway is depth and 3-D shape perception (Welchman, 2016), and it has been suggested that representations in these areas may arise from self-supervised predictive coding (Jehee et al., 2006; Raman & Sarkar, 2016; Bakhtiari et al., 2021).

What kinds of neural networks might best account for neural processing in the parietal visual stream? We propose that the "GRNN" model from Tung et al. (2019) is a promising "proxy model" (Leeds et al., 2013) for investigating computational constraints within the parietal pathway. GRNN learns spatially-aware 3D representations of visual inputs and is trained in a self-supervised manner to predict the complete 3D feature representation of a scene from one camera viewpoint, given input from another camera viewpoint. The model can "fill in" and predict features in the 3D feature map to represent the complete 3D geometry and shape of a scene from a partial 2.5D view. In this sense, the task solved by the GRNN model aligns well with several of the commonly proposed functional characteristics of the parietal stream.

[1] Following Finzi et al. (2022), we refer to the dorsal stream as the parietal stream to avoid confusion with the lateral stream.

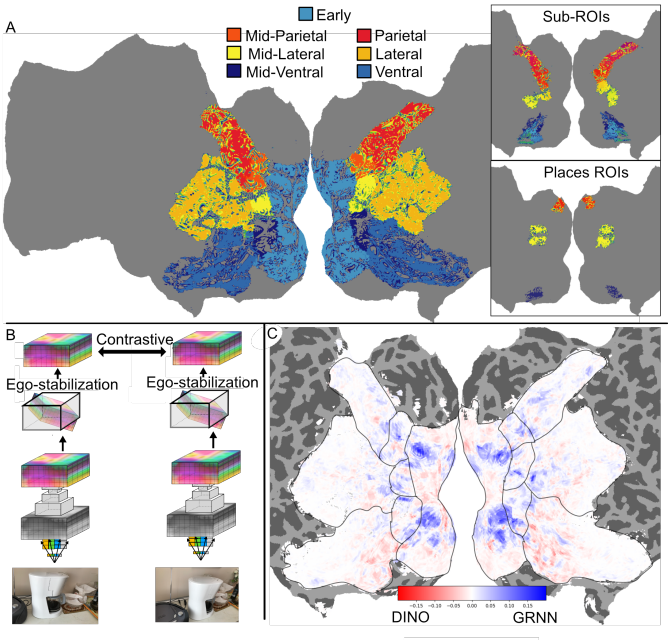

Figure 1: **A.** ROI definitions on a flat map of the fsaverage cortex. **B.** GRNN training procedure. **C.** Difference (subtraction) between GRNN and DINO noise-corrected R^2's for each voxel for example S1. We observe higher R^2 for GRNN (darker blue) in high-level parietal regions, and conversely higher R^2 for DINO (darker red) in high-level ventral regions.

We examined GRNN's ability to predict neural responses as measured by fMRI in response to viewing complex, natural scenes (Allen et al., 2022). As baseline models, we used self-supervised DNNs that were trained to maximize agreement between different augmentations of 2D images, which have previously been shown to be highly predictive of ventral visual stream (Caron et al., 2021; Chen et al., 2020). Consistent with our proposal, we found that the GRNN model was typically a better predictor of high-level parietal visual areas, while the self-supervised 2D models were typically better predictors of high-level ventral visual areas. These results demonstrate the potential for using task-relevant models aligned with hypotheses regarding brain function as a means for probing representational differences across visual streams.

Methods

fMRI dataset. NSD contains measurements of 7T fMRI responses (1.8 mm, 1.6 s) from 8 participants who each viewed 9,000–10,000 distinct color natural scenes (22,000–30,000 trials). Participants fixated centrally and performed a long-term continuous image recognition task. The noise ceiling (NC) was estimated in each voxel as described in Allen et al. (2022). We only include voxels with NC $\geq$ 10% variance and report noise-ceiling normalized prediction accuracy.

Regions of Interest (ROIs). We used NSD's "streams" anatomical atlas to define seven ROIs that cover the parietal, lateral, and ventral visual streams (Fig. 1A; also see Finzi et al. (2022)). We also looked at sub-regions within each stream using Glasser et al. (2016) and Wang et al. (2015) atlases. We also examined three scene ROIs (RSC, OPA, and PPA) obtained by thresholding the category functional localizer.

GRNN training and inference. We used the Fang et al. (2020) dataset of RGB-D images ($n = 28345$) of indoor (Straub et al., 2019) and outdoor (Dosovitskiy et al., 2017) scenes for our GRNN training. The self-supervised training procedure from Harley et al. (2019) was used, which utilizes a view-contrastive loss in feature space. This involves back-projecting an RGB image into a 3D voxel grid, deriving a 3D feature map, and pulling corresponding features together from egomotion-stabilized 3D feature maps (Fig. 1B). To extract GRNN representations for NSD images, we used a fixed camera field of view and estimated depth maps using MiDaS (Ranftl et al., 2020).

Comparison models. We compared GRNN to two self-supervised DNNs that have shown exceptional performance in object recognition and ventral stream predictivity, even rivaling supervised models (Zhuang et al., 2021). These models, DINO (Caron et al., 2021) (ViT-small backbone) and SimCLR SimCLR (Chen et al., 2020) (ResNet-50 backbone), have different self-supervised learning objectives and neural architectures. We trained all models on the same dataset of indoor and outdoor scenes to ensure a fair comparison with GRNN.

Fitting to brain data. We evaluated the performance of each model on a held-out test set using an 85:15 validation split for each subject separately. To reduce dimensionality, we used PCA to project the features into a lower dimensional subspace and retained the first 1000 components (Schrimpf et al. (2018)). We fit the features of each layer to each brain voxel using ridge regression, and determined each voxel's regularization parameter through 7-fold cross-validation. We assessed model performance on the test data using Pearson's correlation and coefficient of determination (R^2), and reported the best fitting layer for each subject in each ROI.

Results

We evaluated the performance of three models for predicting voxel responses to natural images. A representative subject (Fig. 1C) indicates that GRNN outperforms DINO in high-level parietal regions, while DINO performs better in high-level ventral regions. Prediction accuracy (R^2) within each stream across all subjects reveals that GRNN predicts high-level parietal regions better than DINO and SimCLR (GRNN>DINO mid-parietal $p = 0.016$, high-parietal $p = 0.038$; GRNN > SimCLR, high-parietal $p = 0.005$). DINO also performs better than GRNN in high-level ventral regions (paired t-test; $p = 0.0052$). There was no significant difference between GRNN and SimCLR in high-level ventral (paired t-test; $p = 0.98$) and no significant difference between models in the mid- and high-level lateral ROIs. Paired t-tests on sub-ROIs within each stream revealed that GRNN predicts voxel responses with higher accuracy than DINO in V3AB ($p = 0.021$), as well as early IPS regions and SPL compared to DINO (GRNN>DINO, IPS0; $p = 0.017$, IPS1; $p = 0.035$, IPS2; $p = 0.039$, SPL; 0.0078) and SimCLR (GRNN>SimCLR, IPS0; $p = 0.018$, IPS1; $p = 0.010$, IPS2; $p = 0.013$, SPL; 0.015). No difference between models was found in higher-order IPS (GRNN>DINO, IPS3; $p = 0.22$, IPS4; $p = 0.35$, IPS5; $p = 0.058$). In high-level ventral regions DINO had significantly higher predictivity than GRNN in 4/6 regions examined (DINO>GRNN, PHC; 0.023, PHA; 0.0019, VMV; 0.047, VVC; $p = 0.00066$). In scene ROIs DINO significantly outperformed GRNN in two scene regions located more ventrally (DINO>GRNN, PPA; $p = 0.0034$, RSC; 0.00091), whereas GRNN performed better in predicting voxel responses in OPA, which is located more dorsally (GRNN>DINO $p = 0.011$; GRNN>SimCLR $p = 0.037$). These results indicate that a self-supervised model trained for 3D view prediction performs better than models trained to capture augmentation-invariant 2D image statistics in predicting voxel responses in parietal areas. The opposite trend was found in ventral regions, particularly with DINO outperforming GRNN.

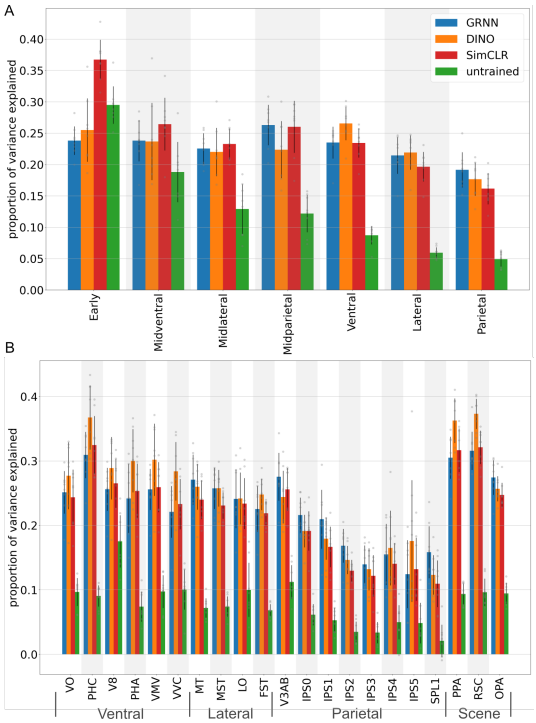

Figure 2: **A.** noise-corrected R^2 of model fit for each ROI. Each dot represents a subject. Bars and error bars are mean and standard deviation across subjects (N=8), respectively. **B.** Same as A but for subregion ROIs.

Discussion

Our research indicates that a 3D view prediction model is better suited for predicting voxel responses in the parietal visual stream compared to 2D augmentation-invariant self-supervised models in a large-scale fMRI dataset of humans viewing natural images. However, more research is necessary to better understand the observed differences and to explore the impact of training and fMRI datasets on model alignment.

Acknowledgments

This material is based upon work supported by National Science Foundation grants GRF DGE1745016 DGE2140739 (GS), a DARPA Young Investigator Award, a NSF CAREER award, an AFOSR Young Investigator Award, and DARPA Machine Common Sense. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Army, the National Science Foundation, or the United States Air Force.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. doi: 10.1038/s41593-021-00962-x

Ayzenberg, V., & Behrmann, M. (2022). The dorsal visual pathway represents object-centered spatial relations for object recognition. *Journal of Neuroscience*, *42*(23), 4693–4710.

Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., & Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, *34*, 25164–25178.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9650–9660).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In *Proceedings of the 1st annual conference on robot learning* (pp. 1–16).

Fang, Z., Jain, A., Sarch, G., Harley, A. W., & Fragkiadaki, K. (2020). Move to see better: Self-improving embodied object detection. *arXiv preprint arXiv:2012.00057*.

Finzi, D., Yamins, D. L., Kay, K., & Grill-Spector, K. (2022). Do deep convolutional neural networks accurately model representations beyond the ventral stream? In *2022 conference on cognitive computational neuroscience.*

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., ... others (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Güçlü, U., & van Gerven, M. A. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, *145*, 329–336.

Harley, A. W., Lakshmikanth, S. K., Li, F., Zhou, X., Tung, H.-Y. F., & Fragkiadaki, K. (2019). Learning from unlabelled videos using contrastive predictive neural 3d mapping. *arXiv preprint arXiv:1906.03764*.

Jehee, J. F., Rothkopf, C., Beck, J. M., & Ballard, D. H. (2006). Learning receptive fields using predictive feedback. *Journal of Physiology-Paris*, *100*(1-3), 125–132.

Leeds, D. D., Seibert, D. A., Pyles, J. A., & Tarr, M. J. (2013). Comparing visual representations across human fMRI and computational vision. *J Vis*, *13*(13), 25. doi: 10.1167/13.13.25

Mineault, P., Bakhtiari, S., Richards, B., & Pack, C. (2021). Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *Advances in Neural Information Processing Systems*, *34*, 28757–28771.

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, *25*(2), 100–110.

Raman, R., & Sarkar, S. (2016). Predictive coding: a possible explanation of filling-in at the blind spot. *PloS one*, *11*(3), e0151194.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, *44*(3), 1623–1637.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... others (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., ... Newcombe, R. (2019). The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*.

Tung, H.-Y. F., Cheng, R., & Fragkiadaki, K. (2019). Learning spatial common sense with geometry-aware recurrent networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2595–2603).

Ungerleider, L. G. (1982). Two cortical visual systems. *Analysis of visual behavior*, 549–586.

Wang, L., Mruczek, R. E., Arcaro, M. J., & Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. *Cerebral cortex*, *25*(10), 3911–3931.

Weiner, K. S., & Grill-Spector, K. (2013). Neural representations of faces and limbs neighbor in human high-level visual cortex: evidence for a new organization principle. *Psychological research*, *77*, 74–97.

Welchman, A. E. (2016). The human brain in depth: how we see in 3d. *Annual review of vision science*, *2*, 345–376.

Wurm, M. F., & Caramazza, A. (2022). Two 'what'pathways for action and object recognition. *Trends in cognitive sciences*, *26*(2), 103–116.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci USA*, *111*(23), 8619–8624. doi: 10.1073/pnas.1403112111

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C.,

DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118.