# Notes on Methodology

**The Pitfall of In-Distribution Evaluation:** A critical finding from Madan et al. (2024) is that DNN-based encoding models, even those with high in-distribution neural predictivity, can break down completely under distributional shifts. The benchmarking showed that models retain as little as ~20% of baseline neural predictivity when tested on images with novel variations in contrast, hue, or saturation. This demonstrates a fundamental failure of generalization. To avoid overinterpreting in-domain performance, claims of brain-alignment must be gated by an explicit OOD check. A rigorous evaluation should involve training and fitting a readout on one image distribution and then testing its neural predictivity on entirely independent OOD image sets (e.g., containing stylizations, sketches, or corruptions) before making strong claims about Brain-Score or IT-predictivity.

**Early visual cortex (V1/V2)**

CNN-style locality and translation equivariance map naturally onto retinotopy and localized receptive fields, and adding an explicit V1-like front end produces causal improvements in V1 predictivity. The VOneNet work is the clearest example: a Gabor-like V1 module put before a standard CNN increases V1 fit and robustness, which shows early-stage inductive biases matter. That said, the gap narrows if models (CNNs or transformers) are given local priors (local attention or hierarchical patching) or hybrid front ends. (Dapello et al., 2020).

**Texture vs. shape (behavioral factorization)**
A well-established finding is that standard ImageNet-trained CNNs exhibit a texture bias, whereas Vision Transformers often show a greater shape sensitivity that is more aligned with human perception. Crucially, this initial architectural prior is strongly modulated by the training regimen. CNNs can be made shape-biased through targeted training on datasets like Stylized-ImageNet, demonstrating that the visual diet can override architectural tendencies. Consequently, large-scale analyses suggest that variation in training data often has a larger and more consistent effect on brain predictivity in high-level visual areas than architectural differences alone.

**Temporal dynamics and recurrence**
Time-resolved signatures in V4/IT often reflect recurrent processing and adaptation that feedforward, single-image models cannot mechanistically implement. Recurrent/dynamical architectures (or shallow recurrent modules added to backbones) typically better capture many time-varying neural responses and behavioral phenomena that depend on temporal evidence accumulation. Depth or engineered readouts can sometimes approximate these effects, but recurrence is a principled way to model dynamics. (Kar et al.; CORnet-style papers).

**Hybrid and modern design trends**
The field is also moving towards hybrid designs (ConvNeXt, CoAtNet, swin-style hierarchies) which combine local feature extraction with long-range attention; these hybrids often preserve early-vision benefits while improving higher-level integration. Large controlled comparisons also show that architecture, scale, data, and training recipe jointly determine brain alignment — no family is a universal winner across all metrics, ie., diverse architectures achieve similar neural predictivity after training and fitting (something observed in experiments). (ConvNeXt / CoAtNet papers; recent controlled studies).

**On forgetting and dataset diversity:** Increased dataset diversity typically improves representation robustness, but sequential fine-tuning on new domains can cause catastrophic forgetting and reduce neural predictivity on earlier distributions. To manage this tradeoff possible practical approach can be to do large-scale pretraining + frozen backbone readouts for neural regressions. Where adaptation is required we will prefer joint multi-dataset training or lightweight adapter modules/LoRA to full fine-tuning; if sequential fine-tuning is unavoidable we will evaluate a standard continual-learning baseline (EWC or rehearsal) and report forgetting metrics explicitly.

**Synthesis / practical implication**
Architecture is a strong prior — it shapes what's easy to learn — but it interacts tightly with the training ecology (data, objective, optimizer, compute). So fair model↔brain comparisons must control or factor these variables (parameter/FLOP matching, optimizer × architecture ablations, multi-seed replication, noise-ceiling reporting, and explicit readout policies) before attributing differences to architecture itself. Large, controlled meta-studies confirm that training and scale frequently explain large portions of variance in brain alignment, so careful deconfounding is essential. Collectively, these findings highlight that model–brain correspondence is not fixed to a single family but emerges from interactions between architectural priors and training ecology.

---

# References


- Dapello, J., Marques, T., Schrimpf, M., Geiger, F., & others. (2020). *Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations (VOneNet).* NeurIPS 2020. NeurIPS Proceedings
- Madan S, Xiao W, Cao M, Pfister H, Livingstone M, Kreiman G. (2024). *Benchmarking out-of-distribution generalization capabilities of DNN-based encoding models for the ventral visual cortex. NeurIPS*

- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.* (Stylized-ImageNet work). arXiv / ICLR discussions. arXiv
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). *An image is worth 16×16 words: Transformers for image recognition at scale.* ICLR 2021 (Vision Transformer). arXiv
- Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). *Evidence that recurrent circuits are critical to the ventral stream's object recognition.* Nature Neuroscience / CBMM report. cbmm.mit.edu
- Conwell, C., et al. (2024). *A large-scale examination of inductive biases shaping high-level visual representation.*Nature Communications. (Controlled comparisons across 224 models.) Nature
- Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). *A ConvNet for the 2020s (ConvNeXt).*CVPR 2022. CVF Open Access
- Dai, Z., et al. (2021). *CoAtNet: Marrying convolution and attention for all data.* NeurIPS 2021. NeurIPS Proceedings
- Cadena, S. A., et al. (2019). *Deep convolutional models improve predictions of macaque V1 responses to natural images.* (Evidence for CNNs matching V1 responses under some conditions.) PMC
- Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). *Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting.* Journal of Cognitive Neuroscience.


---