Play the (Mis)Match: Using fMRI-Aligned Feature Fine-Tuning to Reveal Shortcut Bias in Deep Neural Networks

Yang Chen Lin[1], **Chiayun Lee**[1], **Po-Chih Kuo**[1]

[1]Department of Computer Science, National Tsinghua University
yangchenlin@gapp.nthu.edu.tw, chiayunlee.99@gmail.com, kuopc@cs.nthu.edu.tw

Abstract

Deep neural networks (DNNs) often "cheat" by relying on shortcut objects (e.g., *food⇒kitchen*) rather than holistic spatial layout, undermining out-of-distribution (OOD) robustness. This work serves as a proof-of-concept exploration of whether fMRI alignment can reduce shortcut bias in visual DNNs. We address this issue with **Play the (Mis)Match**, a diagnostic dataset and brain-aligned fine-tuning framework. Leveraging fMRI recordings from the Natural Scenes Dataset (four participants; bedroom, bathroom, living room, kitchen), we curate MATCH images in which shortcut cues co-occur as usual and MISMATCH images from which those cues are removed. ImageNet-initialised CNN and Transformer backbones are fine-tuned with an MSE alignment loss that steers their intermediate features toward voxel patterns known to be less sensitive to shortcut cues. Our results show that, for ResNet, this procedure narrows the Match–Mismatch accuracy gap by 24 % and redirects Grad-CAM attention from individual objects to holistic scene structure, particularly activity from the scene-selective cortex (PPA, RSC, OPA), all without explicit shortcut annotations. Our study provides a proof-of-concept that human-brain constraints may help steer DNNs toward more semantically grounded, less shortcut-dependent scene representations.

Introduction

Understanding how intelligent systems extract meaning from the visual world is a central goal of both cognitive neuroscience and machine learning. Deep neural networks (DNNs) have made impressive strides in visual recognition, often reaching or surpassing human-level performance on benchmark datasets. However, a persistent limitation remains: deep networks frequently rely on *shortcut learning*—the use of unintended, spurious correlations that lead to brittle generalization (Geirhos et al. 2020; Steinmann et al. 2024; Hermann et al. 2023; Taori et al. 2020). Rather than learning to represent the semantic and spatial structure of a scene, DNNs often base predictions on easily accessible but semantically shallow cues such as textures, co-occurring objects, or background patterns (Geirhos et al. 2018; Hermann, Chen, and Kornblith 2020). As a result, their performance collapses under distribution shifts where such shortcuts are no longer valid.

Evidence from cognitive neuroscience indicates that the human brain organizes both concrete and abstract information into structured, low-dimensional representations that support generalization. Scene-selective regions such as the parahippocampal place area (PPA), the retrosplenial complex (RSC) and the occipital place area (OPA) play a key role in encoding global scene layout, semantic context and prior knowledge, rather than isolated objects (Peer et al. 2021).

In this paper, we curate **Play the (Mis)Match** (Figure 1), a dataset to probe what objects are shortcuts for DNNs to recognize scenes, focusing on four indoor scene categories: Living Room, Bedroom, Kitchen, Bathroom, and Outdoor scenes. Each category was divided into two test subsets: **Match** (i.i.d.), in which shortcut objects (e.g., TV in living rooms, toilets in bathrooms) co-occurred as expected, and **Mismatch** (o.o.d.), in which those shortcut objects were explicitly absent. This allows us to directly probe models' reliance on shortcut features and their ability to generalize in their absence in downstream scene classification tasks. The core idea of our approach is to leverage functional magnetic resonance imaging (fMRI) signals from the human brain to guide deep models away from shortcut features and toward more spatially grounded, potentially more generalizable representations. In particular, we fine-tune DNNs using fMRI-derived activation patterns from OPA, PPA, RSC, and other visual regions, selecting only those voxels that reliably distinguish scene categories, independent of shortcut-object presence. Our training protocol incorporates these brain signals as auxiliary targets, encouraging the model to align its internal representations with those of the human brain under conditions where shortcut features are withheld. Our work makes the following contributions:

- We introduce a brain-informed training framework for mitigating shortcut bias in DNN by aligning model representations with fMRI responses insensitive to shortcuts.
- We curate a dataset with match (i.i.d.) and mismatch (o.o.d.) scene categories for directly testing shortcut object dependence in scene classification tasks.
- We present initial evidence that fMRI-aligned fine-tuning can modestly reduce shortcut reliance and improve generalization tendencies, even without shortcut annotations.

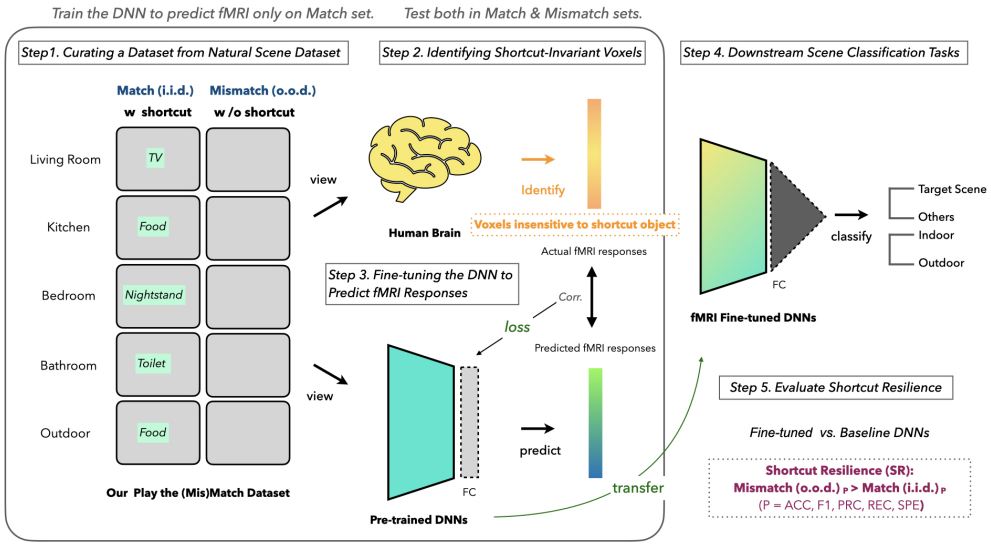


Figure 1: Overview of Play the (Mis)Match. Two distribution scene datasets with (i.i.d.) and without (o.o.d.) shortcut objects. Each model is fine-tuned to predict voxel-wise fMRI responses from image inputs. Predicted fMRI responses guide the model away from shortcut object features and toward distributed, biologically plausible scene representations. The procedure evaluates generalization by comparing performance on downstream scene classification tasks across match and mismatch sets.

What is Shortcut Learning?

Geirhos et al. (Geirhos et al. 2020) define shortcut learning as the tendency of machine learning models to adopt *decision rules that perform well on standard benchmarks*—typically i.i.d. test data—but fail to generalize under distribution shifts or more challenging testing conditions. This discrepancy reveals a fundamental mismatch between the *intended solution* and the one actually learned by the model. This behavior has been observed across domains, including image classification, medical diagnosis, sentiment analysis, and multilingual translation. Formally, let:

- f_θ denote a model learned with parameter θ be trained on dataset D_{train}.
- $D_{\text{i.i.d.}}$ represents in-distribution test data, sampled from the same distribution as D_{train}.
- $D_{\text{o.o.d.}}$ represent out-of-distribution data with altered feature correlations.
- $\mathcal{P}(f_\theta, D)$ be the performance of f_θ on dataset D.
- F_{intended} denotes the set of task-relevant features (e.g., spatial layout in scene classification).
- F_{shortcut} denotes the set of unintended but predictive features (e.g., presence of food in indoor scenes).

Shortcut learning is considered a fundamental property of overparameterized, nonlinear models trained via empirical risk minimization, where the loss function incentivizes finding any signal (*shortcut opportunities*) that reduces error—even if it is fragile or semantically meaningless (Hermann et al. 2023; Steinmann et al. 2024).

Formally, given a training distribution $\mathcal{D}_{\text{train}}$, a model is optimized to minimize empirical risk:

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}}[\mathcal{L}(f_\theta(x), y)], \tag{1}$$

yet without explicit constraints on feature selection, the model may converge on *shortcut features*—attributes that are easier to extract but not causally or semantically related to the target y. These opportunities manifest as statistical regularities inherent in real-world distributions (e.g., cows typically appearing on grass), dataset-specific biases that create artificial co-occurrence priors, and historical or social biases embedded in the underlying data distributions. Such regularities enable models to achieve high performance on i.i.d. test sets but fail to generalize under distribution shifts in which these spurious correlations no longer hold. Thus, shortcut learning depends on both the feature structure in the data and the model's *decision rule*, i.e., how it selects and combines features during training. This underscores the importance of enforcing inductive biases or alignment strategies that can steer models toward more generalizable solutions.

To better characterize this tendency, Hermann et al. (Hermann et al. 2023) introduce two key notions: *predictivity* (ρ), which measures how informative a feature is with respect to the task, and *availability* (α), which captures how easily a feature can be extracted by a model. Given two latent features z_c (core) and z_s (shortcut) We denote latent feature representations $z_c \in F_{\text{intended}}$ and $z_s \in F_{\text{shortcut}}$, with associated predictivity ρ and availability α.

$$\begin{aligned} \text{reliance}_{\text{model}}(z_s) &= \frac{\alpha_s\,\rho_s}{\alpha_c\,\rho_c + \alpha_s\,\rho_s}, \\ \text{reliance}_{\text{optimal}}(z_s) &= \frac{\rho_s}{\rho_c + \rho_s}, \end{aligned} \tag{2}$$

$$\text{bias} = \text{reliance}_{\text{model}}(z_s) - \text{reliance}_{\text{optimal}}(z_s) > 0. \tag{3}$$

Hence, the model exhibits a positive bias towards when the shortcut feature z_s are substantially more available ($\alpha_s \gg \alpha_c$), the available predictive signal for z_s outweighs that for z_c, even if less predictive ($\rho_s < \rho_c$). As a result, the model "cheats" by exploiting z_s, rather than learning robust, semantically aligned representations through z_c. Consequently, when evaluated on $D_{\text{o.o.d.}}$, where shortcut object correlations are deliberately disrupted, the model's performance $\mathcal{P}(f_\theta, D_{\text{o.o.d.}})$ deteriorates significantly.

Related work

Detecting and mitigating Strategies for Shortcuts

Based on the taxonomy proposed by (Steinmann et al. 2024), existing **shortcut detection methods** fall into four broad categories: (1) *utility-based* approaches that monitor differential learning dynamics, such as mutual information tracking (Adnan et al. 2022) or simplicity bias (Yang et al. 2023); (2) *perturbation-based* diagnostics using frequency filtering (Chormai et al. 2024) or adversarial examples (Geirhos et al. 2020); (3) *interpretability tools* that visualize model reliance on specific input features (Carter et al. 2021); and (4) *causal interventions*, which rely on interventional or counterfactual data (Zheng and Makar 2022). **Shortcut mitigation strategies** operate at three levels: *data*, *model*, and *inference*. At the data level, approaches include targeted dataset editing or rebalancing to decorrelate spurious cues from labels (He, Shen, and Cui 2021). Model-level techniques include architectural changes, loss regularization, and orthogonalization of latent features (Steinmann et al. 2024).

While effective in constrained domains, existing shortcut detection/mitigation techniques predominantly rely on explicit supervision signals or prior knowledge regarding spurious feature distributions—assumptions that prove impractical in real-world scenarios and brittle under distribution shift. In contrast, our approach leverages biologically meaningful latent representations by training models to predict fMRI response patterns associated with high-level scene categorization—specifically independent of shortcut object presence—thereby introducing an inductive bias that circumvents the need for explicitly labeled spurious features. Additionally, our curated dataset provides a standardized benchmark for quantitative evaluation of o.o.d. generalization capabilities (how well the model's resilience to shortcuts), facilitating systematic comparison across different model behaviors on various tasks.

Human Model Alignment

Vision models fine-tuned on human behavioral responses, including similarity judgments (Sundaram et al. 2024; Muttenthaler et al. 2023) and eye-gaze data (Lopez-Cardona et al. 2024), demonstrate improved performance across segmentation, retrieval, and reward prediction tasks. A growing body of work aims to bridge the representational gap between DNNs and the human brain (Sucholutsky et al. 2023; Xu and Vaziri-Pashkam 2021). Recent studies emphasize that *mid-level perceptual features*—such as spatial layout, object pose, and scene geometry—are crucial for achieving human-like generalization (Opiełka, Loke, and Scholte 2024; St-Yves et al. 2023). (Ren and Bashivan 2024) evaluated 38 DNNs on their ability to predict neuronal responses in primate area V4 under i.i.d. and o.o.d. conditions. While models performed well on naturalistic images, they exhibited large generalization gaps on synthetic stimuli (e.g., sketches). Crucially, standard object recognition benchmarks (e.g., ImageNet accuracy) did not predict neural alignment. Instead, *adversarial robustness* and model properties such as depth and ensemble diversity were stronger predictors of o.o.d neural predictivity—suggesting that robustness-oriented training leads to more brain-like representations. Other large-scale studies (Conwell et al. 2024) further reveal that architecture and training objective alone do not determine brain alignment. A comprehensive evaluation of 224 pre-trained DNNs against fMRI responses from the Natural Scenes Dataset (NSD) showed that the *diversity and richness of visual data experience*—rather than architectural class (e.g., CNN vs. ViT) or supervision type (e.g., contrastive vs. caption-based)—was the dominant factor influencing brain predictivity. Complementary evidence from the THINGS neuroimaging dataset (Hebart et al. 2023), which supports a multidimensional view of object representation in the brain. Researchers identified 66 interpretable object dimensions, spanning perceptual and semantic axes, and showed that fine-tuned CLIP-ViT models could predict these dimensions and explain cortical variance beyond categorical models (Contier, Baker, and Hebart 2024).

Together, these studies underscore that brain-like models must go beyond optimizing for task performance (Feather et al. 2023; Wichmann and Geirhos 2023; Kubilius et al. 2019). They must also incorporate structural inductive biases and objective functions that reflect the brain's principles of abstraction, sparsity, and compositionality. Encoding models predict voxel-wise brain responses from inputs or DNN features (Matsuyama, Sasaki, and Nishimoto 2023; Xu and Vaziri-Pashkam 2021; Muttenthaler et al. 2024). Our method reverses this: our work contributes to this direction by using fMRI responses (from NSD) to guide DNN training: fine-tuning vision models to predict fMRI responses (Adeli, Minni, and Kriegeskorte 2023; Yang, Gee, and Shi 2024), specifically selecting voxels that capture scene category distinctions independently of shortcut object presence. This biologically constrained alignment reduces shortcut reliance and improves model generalization under distribution shift. Moreover, our experiments reveal and quantify shortcut features embedded in standard scene datasets, offering a principled framework for diagnosing and mitigating shortcut bias.

Method

Step 1: Curating a Shortcut Dataset

Natural Scenes Dataset (NSD). We leverage the NSD (Allen et al. 2022), a large-scale neuroimaging corpus com-

prising 7T functional MRI responses from eight subjects viewing 10,000 unique naturalistic images primarily sampled from MS-COCO. Each stimulus was presented multiple times across scanning sessions, enabling the extraction of trial-averaged BOLD response patterns with improved signal-to-noise. While the complete NSD comprises data from eight participants, we restrict our analyses to four subjects (1, 2, 5, 7) due to computational constraints and data-quality considerations. The preprocessed NSD data provide reconstructions of the cortical surface for each subject, where each vertex corresponds to a spatial location in the visual cortex, yielding high-dimensional activation vectors $\mathbf{v}_i \in \mathbb{R}^d$ for each stimulus presentation.

Scene Categories and Shortcut Objects. We define a taxonomy of scene categories $\mathcal{C} = \{\text{Li}, \text{Ki}, \text{Be}, \text{Ba}, \text{O}\}$ corresponding to living room, kitchen, bedroom, bathroom and outdoor environments, respectively (Figure 2). For each scene category $y_c \in \mathcal{C}$, we identify a diagnostic shortcut object $o_c \in \mathcal{O}$ based on co-occurrence statistics from COCO metadata (Lin et al. 2014). Specifically, we define the mapping:

$$\mathcal{M}_{\text{obj}} : \mathcal{C} \to \mathcal{O} \tag{4}$$

by $\mathcal{M}_{\text{obj}}(\text{Li}) = \text{TV}, \mathcal{M}_{\text{obj}}(\text{Ki}) = \text{food}, \mathcal{M}_{\text{obj}}(\text{Be}) = \text{nightstand}, \mathcal{M}_{\text{obj}}(\text{Ba}) = \text{toilet}, \mathcal{M}_{\text{obj}}(\text{O}) = \text{food}$. We then construct two data distributions:

$$\begin{aligned} \mathcal{D}^{\text{Match}} &= \big\{(x,\, y,\, o) \mid y \in \mathcal{C},\ o = \mathcal{M}_{\text{obj}}(y)\big\}, \\ \mathcal{D}^{\text{Mismatch}} &= \big\{(x,\, y,\, o) \mid y \in \mathcal{C},\ o \neq \mathcal{M}_{\text{obj}}(y)\big\}. \end{aligned} \tag{5}$$

In $\mathcal{D}^{\text{Match}}$, where each scene x contains its canonical shortcut object o associated with category y; in $\mathcal{D}^{\text{Mismatch}}$, where object-scene associations are deliberately violated to break shortcut reliance.

Scene Classification Task.

- Task1: Scene-Specific Indoor Classification. For each target category $y_{\text{target}} \in \{\text{Li}, \text{Ki}, \text{Be}, \text{Ba}\}$, we train a binary classifier of one vs. rest that distinguishes images of y_{target} from all other indoor scenes.
- Task2: Indoor–Outdoor Classification. We train a binary classifier to discriminate all indoor scenes $(\text{Li}, \text{Ki}, \text{Be}, \text{Ba})$ versus outdoor.

Step 2: Identifying Shortcut-Invariant Voxels

To obtain biologically meaningful regularization targets, we applied contrast-based statistical masking to remove shortcut-related activations and retain voxels encoding genuine scene-level semantics. The resulting *Shortcut-Invariant Voxels* serve as fMRI-derived supervision signals for fine-tuning deep networks toward spatially grounded, context-sensitive representations.

Scene-Specific Indoor Classification. We implemented a hierarchical masking framework to isolate neural populations that encode scene layout rather than shortcut objects. For each indoor target category, we computed a series of general linear model contrasts on trial-averaged BOLD responses (FDR-corrected, $p < 0.05$). Voxels were sampled from scene-selective regions (PPA, RSC, OPA) and, for comparison, from early visual areas (V1–V4) and object-selective regions.

Two types of contrasts were generated: (1) *Between-category* contrasts distinguishing each target scene from all other indoor scenes (e.g., Living Room vs. Kitchen, Bathroom, Bedroom); and (2) *Within-category* contrasts capturing object sensitivity by comparing images with vs. without the shortcut object.

We then applied statistical masking to subtract object-sensitive activations from the scene-discriminative maps, yielding voxel subsets responsive to spatial semantics but invariant to diagnostic objects:

$$\begin{aligned} \text{Shortcut-Invariant Voxel}_{\text{Li}} = (\text{Voxel}_{\text{Li,Ki}} \cap \text{Voxel}_{\text{Li,Ba}} \\ \cap \text{Voxel}_{\text{Li,Be}}) - \text{Voxel}_{\text{Li}_1,\text{Li}_2}. \end{aligned} \tag{6}$$

Indoor–Outdoor Classification. For the indoor–outdoor paradigm, we constructed four voxel-wise statistical maps (FDR-corrected, $p < 0.05$): Voxel_a capturing category-selective responses when the food shortcut was present; Voxel_b capturing responses when it was absent; Voxel_c isolating food-related activations within indoor contexts; and Voxel_d identifying context-inappropriate object responses in outdoor scenes.

A masking operation was then applied to retain consistent category-selective activations while excluding shortcut-driven voxels:

$$\begin{aligned} \text{Shortcut-Invariant Voxel}_{\text{Indoor–Outdoor}} = (\text{Voxel}_a \cap \text{Voxel}_b) \\ - (\text{Voxel}_c \cup \text{Voxel}_d). \end{aligned} \tag{7}$$

Although this masking procedure relies on contrasts defined with respect to known shortcut objects (e.g., TV, toilet, food), this step serves as an *experimental control* rather than a form of supervision leakage. Following standard cognitive neuroscience practice, we treat object-presence contrasts as controlled confounds that allow us to *isolate* rather than encode shortcut sensitivity. This ensures that the subsequent alignment procedure (Step 3) trains DNNs to predict voxel responses that genuinely reflect scene-level semantics rather than shortcut-driven activations, providing a biologically interpretable supervision signal for model fine-tuning.

Step 3: Fine-tuning DNN to predict fMRI responses

Summary of Training Data. Our training protocol leveraged a stratified data set comprising neuroimaging data from four subjects (1, 2, 5, 7) in two distinct classification paradigms. For task 1 (scene-specific indoor classification), we implemented a one-versus-rest framework in four indoor scene categories. The dataset exhibits inherent class imbalance, with target categories (Living room, Bathroom, Kitchen, Bedroom) represented by 63-157, 92-135, 91-138, and 79-108 images, respectively, across subjects. Task 2 (indoor-outdoor classification) employed a binary discrimination paradigm with more balanced class distribution, incorporating 92-153 indoor and 87-134 outdoor

Figure 2: Examples from the **Play the (Mis)Match** dataset.

exemplars per subject. Living room discriminative voxels ranged from 889 (Subject 7) to 2814 (Subject 2), while bathroom-selective patterns exhibited the highest dimensionality (4265-8295 voxels). The kitchen and bedroom categories showed comparatively constrained neural representations (289-875 and 584-2798 voxels, respectively). For coarser indoor-outdoor discrimination, the number of selected voxels ranged from 570 (Subject 7) to 1790 (Subject 1), reflecting the spatial extent of category-selective activation patterns within the visual processing hierarchy. This subject-specific dimensionality variance underscores the importance of individualized neural alignment strategies in our regularization framework. In total, the training set comprised 9,842 images, and the test set contained 160 held-out examples.

Training Objective. To steer models away from object-centric shortcuts and toward robust, context-sensitive scene representations, we fine-tune the DNNs to predict voxel-level fMRI responses from scene-selective regions (PPA, RSC, OPA) and, for comparison, from early visual areas used as a baseline (V1–V4) and object-selective regions. By focusing on voxels that reliably distinguish between scene categories, this process encourages models to internalize global and semantically grounded features aligned with human visual perception. We begin with pretrained CNNs and Transformer-based models on ImageNet (Krizhevsky, Sutskever, and Hinton 2012) as baseline (B) models, including ResNet (He et al. 2016), VGGNet (Simonyan and Zisserman 2014), EfficientNet (Tan 2019), MobileNetV2 (Sandler et al. 2018), DenseNet (Huang et al. 2017), InceptionV3 (Szegedy et al. 2016), Vision Transformer (ViT) (Dosovitskiy 2020), and DINOv2 (Oquab et al. 2023). For each backbone, we attach a lightweight fully connected layer g_θ to the penultimate features. Given an input image x, the image encoder produces mid-level activations $f_\theta(x)$; the readout maps them into fMRI space, $\hat{v} = g_\theta\big(f_\theta(x)\big) \in \mathbb{R}^d$, where d is the number of voxels retained after selection. For each image x sampled from either $\mathcal{D}^{\text{Match}}$ or $\mathcal{D}^{\text{Mismatch}}$, we obtain its measured voxel-response vectors $\mathbf{resp}_{\text{Mismatch}}$ or $\mathbf{resp}_{\text{Match}} \in \mathbb{R}^d$. We first retain only those voxels that are *insensitive* to shortcut objects (see Step 2 Identifying Shortcut-Invariant Voxels). This contrast $\Delta\mathbf{resp}$ preserves activity driven by scene layout (F_{intended}) while suppressing shortcut-object responses (F_{shortcut}). We then train the network to predict $\Delta\mathbf{resp}$ with a mean-squared-error (MSE) loss:

$$\mathcal{L}_{\text{fMRI}-\text{aligment}}(\theta) = \frac{1}{d}\left\|\hat{v} - \Delta\mathbf{resp}\right\|_2^2. \tag{8}$$

This fMRI-alignment loss serves as a biologically inspired regularizer, guiding the model to produce internal representations that are more consistent with those (scene–level features) encoded in human cortical responses, while ignoring shortcut cues. The target Δresp corresponds to the residualized voxel response vectors obtained after Step 2. The resulting alignment supports generalization under distribution shift by reducing reliance on shortcut features during downstream classification. Our fMRI-alignment regularizer encourages f_θ to capture F_{intended} rather than F_{shortcut} (see Eq. (3)). By minimizing the alignment loss $\mathcal{L}_{\text{fMRI-alignment}}$, we expect

$$\text{reliance}_{\text{model}}(z_s) \downarrow \quad \text{and} \quad \text{reliance}_{\text{optimal}}(z_s) \uparrow,$$

thereby shrinking the performance gap impacted on $D_{\text{o.o.d.}}$.

Step 4: Scene Classification

After the DNN is fine-tuned in fMRI prediction, we transfer the fine-tuned (F) DNN to scene classification tasks. Specifically, we attach a standard classification head (a fully connected layer) to the top of the fine-tuned backbone and train it to distinguish among scene categories.

Evaluation Metrics. Following scene classification, we evaluated the model's capacity to generalize beyond shortcut dependencies through a systematic transfer learning protocol. Specifically, we augment each fine-tuned backbone with a task-specific classification head (a fully connected layer with a softmax activation) and train it to discriminate between scene categories using standard cross-entropy loss. To quantify the efficacy of our fMRI-aligned regularization approach, we employ a distributional shift evaluation paradigm. We formalize our primary evaluation metric, **Shortcut Resilience (SR)**, as:

$$\text{SR} = \mathcal{P}(f_\theta, \mathcal{D}^{\text{mismatch (o.o.d.)}}) - \mathcal{P}(f_\theta, \mathcal{D}^{\text{match (i.i.d.)}}), \tag{9}$$

where f_θ represents the model with parameters θ and ($\mathcal{P} \in \{\text{Precision}, \text{Recall}, \text{F}_1, \text{Accuracy (ACC)}, \text{Specificity}\}$ denotes the performance metric. Each performance metric is

presented with its mean value and standard deviation across subjects. **Lower values of SR (Performance gap, ΔP) indicate reduced shortcut features reliance and exhibit enhanced resilience when faced with distribution shifts**, with $SR \approx 0$ suggesting that the model has developed invariance to the presence/absence of shortcut objects and instead uses robust features of the scene for classification. This comprehensive evaluation framework enables precise quantification of whether fMRI-guided regularization induces models to learn context-sensitive, layout-aware representations that generalize effectively when diagnostic shortcuts are removed from the input distribution. To describe how well the brain and model align in encoding, we incorporate Representational Similarity Analysis (Kriegeskorte, Mur, and Bandettini 2008). Given a batch of N stimuli $\{x_i\}_{i=1}^{N}$, we compute representational dissimilarity matrices (RDMs) for both model and brain responses. We then computed the Spearman correlation between the vectorized RDMs.

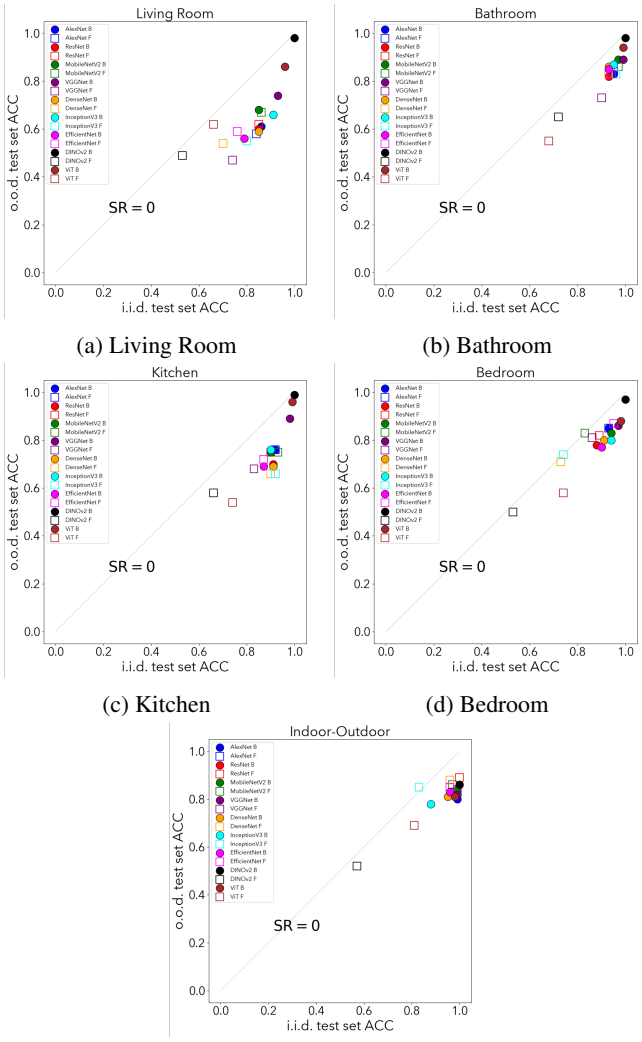
Results

Model Performance and Behavior Across Task

Figure 3 quantifies each model's robustness across scene-specific indoor classification tasks and indoor-outdoor discrimination. Under i.i.d. conditions (x-axis), all models (circles: baseline, squares: fine-tuned) demonstrate high accuracy (0.85), with transformers achieving superior performance (0.95–0.98). Distribution shifts induce architecture-dependent degradation patterns: CNNs exhibit substantial performance decrements (0.25 ACC) on the Living Room and Bathroom tasks, whereas transformers exhibit attenuated decrements (0.10–0.15 ACC). Fine-tuned models predominantly exhibit a rightward-upward trajectory in ACC-space, with intermediate-depth CNNs (ResNet, DenseNet, InceptionV3) demonstrating the most substantial gains (≈0.08–0.10 ACC), approaching the diagonal line ($\Delta_{SR} = 0$) in kitchen and bedroom tasks. ResNet (highlighted in red) exhibits consistent rightward-upward displacement and closer to a diagonal line across all tasks, with a corresponding significant decrease in averaged shortcut reliance (improving from -0.162 to -0.124, $p < .05$) (Figure 4), indicating reduced susceptibility to spurious feature correlations. Conversely, fine-tuned VGGNet demonstrates diminished o.o.d. resilience with downward-leftward displacement. MobileNetV2 and EfficientNet display near-complete recovery from distribution shift effects in indoor-outdoor task.

Model's Attention Shifts

Figure 5 shows the ResNet attention across all five scene categories. fMRI-aligned fine-tuning results in a striking reallocation of model attention (see the red circle). In the Living Room, the baseline model repeatedly latches onto a single salient object (the television); after fine-tuning, attention fans out across sofas, tables, and secondary electronics, yielding a more distributed, "ensemble" representation of the space. In the Bathroom, the unaligned network's heatmap collapses onto the toilet bowl, whereas the fine-tuned model suppresses that cue and instead attends to sur-



(a) Living Room (b) Bathroom

(c) Kitchen (d) Bedroom

(e) Indoor vs. Outdoor

Figure 3: Classification performance across Match (i.i.d.) and Mismatch (o.o.d.) test sets for each model.

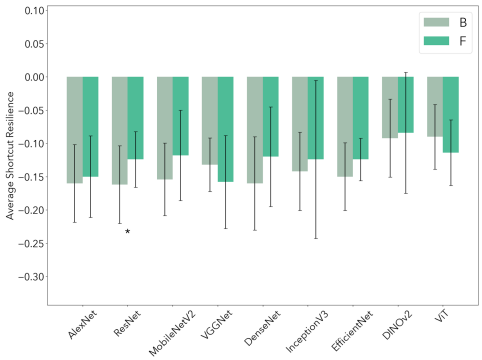


Figure 4: Averaged Shortcut Resilience (SR) across tasks by model. Baseline (B), Finetuned (F).

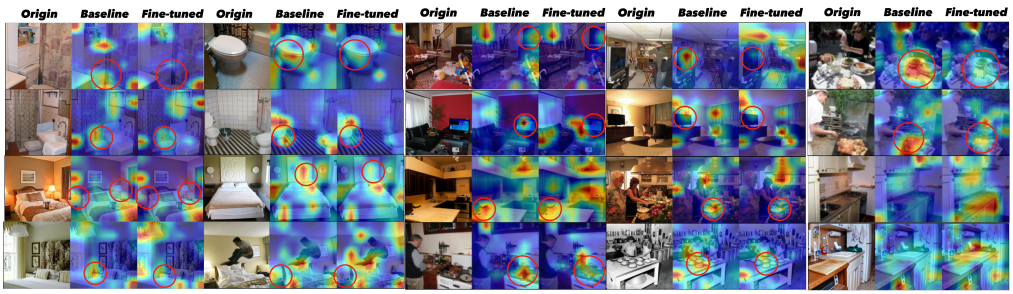


Figure 5: Grad-CAM visualizations comparing RestNet attention before and after fMRI-based fine-tuning

rounding fixtures and spatial context. In the Kitchen, baseline activations often highlight food items as spurious shortcuts, whereas fine-tuned models attenuate these signals. The o.o.d. kitchen showed that the fine-tuned models focused on functional preparation surfaces—counters, stovetops, and cookware (compare to the baseline model). In the Bedroom, vanilla models fixate on decorative artifacts such as lamps and nightstands, whereas fine-tuned variants suppress those distractions and concentrate on structural elements—the bed frame and headboard geometry. Together, these Grad-CAM results suggest that fMRI-guided fine-tuning may help shift model attention away from highly available but spurious cues, potentially improving out-of-distribution generalization.

Discussion

Prior work has demonstrated that incorporating early-vision constraints can enhance the robustness of convolutional networks. For example, through neural regularization in mouse V1 (Li et al. 2019), fixed V1-inspired front-ends in VOneNets (Dapello et al. 2020), and ventral-stream shape bias captured by response-optimized models(Khosla et al. 2022). Extending this idea beyond low-level vision, our work explores whether alignment with mid-level, scene-selective cortical regions (e.g., PPA, RSC, OPA) can similarly influence network behavior. **This proof-of-concept** study provides preliminary evidence that fMRI-based *representational alignment* could introduce an inductive bias that mitigates shortcut bias in principle. By aligning model features with neural activity in scene-selective cortical regions (e.g., PPA, RSC, OPA), our approach encourages networks to attend to spatially distributed and semantically relevant features rather than overfitting to salient shortcut objects. Importantly, the MATCH–MISMATCH diagnostic dataset played a critical role in revealing these effects: by explicitly controlling for shortcut-object co-occurrence, it enabled precise evaluation of how models behave under feature perturbations, highlighting not only classification outcomes but also their reliance on spurious versus holistic cues.

Previous approaches typically try to remove or rebalance shortcut features *in the data* (by editing or reweighting samples) or *in the loss function* (by penalizing spurious correlations). Our approach does something different: instead of changing the data or the loss, we change what the model is encouraged to *see*. By training the network to predict brain activity patterns from scene-selective areas, we nudge it to represent scenes more closely to how humans perceive spatial layout. This means that the model learns to rely less on easy visual shortcuts, such as "there is a TV, so this must be a living room," and more on the overall structure of the scene. In other words, we use the brain signal as a teacher to improve visual representation, not as an additional label or dataset filter.

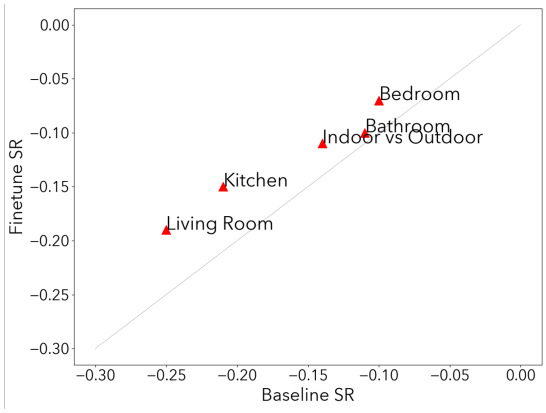


Figure 6: Finetuned vs. baseline ResNet's SR by tasks.

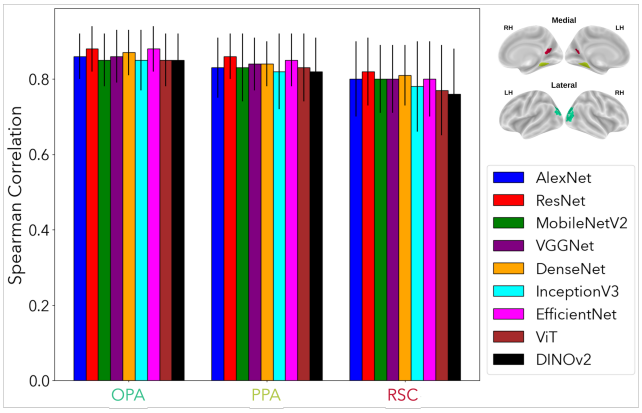


Figure 7: Brain-model alignment.

Among all tested architectures, ResNet showed the clearest trend toward improved shortcut following fMRI-guided

Table 1: Prediction of the corresponding image class using responses from different brain regions

	Indoor-Outdoor		Living Room		Bathroom		Kitchen		Bedroom	
	i.i.d.	o.o.d.	i.i.d.	o.o.d.	i.i.d.	o.o.d.	i.i.d.	o.o.d.	i.i.d.	o.o.d.
V1	0.53 ±0.09	0.51 ±0.04	0.58 ±0.06	0.52 ±0.06	0.66 ±0.05	0.60 ±0.04	0.66 ±0.08	0.57 ±0.05	0.56 ±0.09	0.53 ±0.06
V2	0.55 ±0.13	0.57 ±0.03	0.60 ±0.04	0.53 ±0.05	0.67 ±0.06	0.56 ±0.03	0.58 ±0.11	0.55 ±0.03	0.67 ±0.07	0.60 ±0.05
V3	0.70 ±0.11	0.60 ±0.05	0.61 ±0.06	0.51 ±0.03	0.65 ±0.04	0.57 ±0.06	0.63 ±0.10	0.56 ±0.04	0.62 ±0.07	0.56 ±0.04
V4	0.69 ±0.08	0.52 ±0.04	0.55 ±0.15	0.46 ±0.07	0.64 ±0.05	0.53 ±0.04	0.62 ±0.06	0.56 ±0.01	0.48 ±0.07	0.46 ±0.02
OPA	0.71 ±0.12	0.67 ±0.03	0.60 ±0.06	0.51 ±0.05	0.65 ±0.04	0.56 ±0.08	0.71 ±0.10	0.60 ±0.04	0.59 ±0.12	0.61 ±0.14
PPA	0.73 ±0.09	0.55 ±0.04	0.60 ±0.06	0.50 ±0.06	0.66 ±0.05	0.52 ±0.04	0.71 ±0.08	0.57 ±0.05	0.50 ±0.09	0.54 ±0.06
RSC	0.77 ±0.10	0.66 ±0.11	0.63 ±0.05	0.54 ±0.05	0.58 ±0.06	0.44 ±0.15	0.73 ±0.08	0.54 ±0.04	0.64 ±0.03	0.61 ±0.04

fine-tuning (Figure 4). Our empirical SR metric (Eq. (9)) is the observable counterpart of the theoretical bias term: when SR ≈ 0, the empirical gap between Mismatch and Match accuracies vanishes, indicating no bias (Eq. (3)). Our results suggest that the fine-tuned ResNet exhibited greater stability across tasks (see Figure 6), with indications of reduced shortcut dependence.

This improvement appears to arise from a combination of architectural advantages and greater representational alignment with the human visual system. As shown in Table 1, a fine-tuned ResNet predicts image classes more accurately when signals are drawn from scene-selective regions (OPA, PPA, RSC) than from early visual areas (V1,V2), and the advantage persists under o.o.d. evaluation, consistent with the hypothesis that fMRI-guided fine-tuning may encourage reliance on higher-level scene representations rather than low-level shortcut cues. Structurally, ResNet's residual connections enable deeper integration of multi-scale features (He et al. 2016; Xu and Vaziri-Pashkam 2021), facilitating richer mid-level representations. Supporting this interpretation, our Post-hoc alignment analysis (see Figure 7) revealed that ResNet achieved the highest Spearman correlations with fMRI responses in PPA (0.86), RSC (0.82). These suggest that fine-tuned ResNet begins to capture image features with biologically grounded scene semantics in latent fMRI space. Moreover, a fine-tuned ResNet predicts image classes more accurately when signals are drawn from scene-selective regions (OPA, PPA, RSC) than from early visual areas (V1,V2), and the advantage persists under o.o.d. evaluation (Table 1). It is consistent with the hypothesis that fMRI-guided fine-tuning may encourage reliance on higher-level scene representations rather than low-level shortcut cues.

However, DINOv2 behaves differently. Standard ERM (Eq. 1) rewards any feature that minimizes training loss, regardless of biological plausibility. Adding the fMRI-alignment term $\mathcal{L}_{\text{fMRI-alignment}}$ constrains the solution space to *neuro-plausible* functions, effectively shrinking the shortcut component $\alpha_s \rho_s$ in Eqs. 2–3. For DINOv2, however, highly predictive and highly available object cues already dominate (large α_s, ρ_s). The alignment loss suppresses shortcut-driven features but cannot immediately increase scene-core predictivity ρ_c, leading to an overall accuracy drop of roughly 50%. This outcome illustrates a natural trade-off: when shortcut features are deeply embedded in the representation, enforcing biological alignment reduces bias but may transiently penalize discriminative power.

Conclusion and Limitation

This study aims as a *feasibility* test: a proof-of-concept exploration of **whether fMRI alignment can reduce shortcut bias**. To isolate this question, we deliberately compared a base model with its fine-tuned counterpart on a single, carefully controlled dataset, rather than running an exhaustive factorial ablation. Unlike most shortcut-mitigation methods that fix the data or the loss, our approach changes what the model learns to represent. We teach the network to predict fMRI responses from scene-selective brain regions, guiding it to perceive scenes more like humans do—based on spatial layout rather than shortcut objects.

This clarity comes at a cost: the dataset contains only a few hundred samples per task, spans a narrow range of scene categories, and supports only a limited set of evaluation metrics. Additionally, our current analysis examines pre- and post-training snapshots using Grad-CAM and SR scores. A fuller picture will require *in-training* probes that continuously track (i) which image-level features are amplified or suppressed and (ii) how those shortcut cues are represented within specific neural latent spaces. Recent progress in representation monitoring (e.g., (Adnan et al. 2022)) offers a springboard for the community to develop such tools. These diagnostics could help identify *when* and *where* shortcut features begin to be pruned, informing strategies that target training models with less bias. Our findings provide initial evidence that such alignment may mitigate shortcut bias in controlled settings. Future work should also assess whether our approach generalizes across neural modalities and model tasks.

Taken together, our findings chart a promising research avenue at the intersection of cognitive neuroscience and ro-

bust machine learning: brain-informed training constraints may help steer DNNs toward more human-aligned inductive biases, providing a principled alternative to conventional shortcut-mitigation methods.

References

Adeli, H.; Minni, S.; and Kriegeskorte, N. 2023. Predicting brain activity using Transformers. *bioRxiv*, 2023–08.

Adnan, M.; Ioannou, Y.; Tsai, C.-Y.; Galloway, A.; and Taylor, G. W. 2022. Monitoring shortcut learning using mutual information. *arXiv preprint arXiv:2206.13034*.

Allen, E. J.; St-Yves, G.; Wu, Y.; Breedlove, J. L.; Prince, J. S.; Dowdle, L. T.; Nau, M.; Caron, B.; Pestilli, F.; Charest, I.; et al. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1): 116–126.

Carter, B.; Jain, S.; Mueller, J. W.; and Gifford, D. 2021. Overinterpretation reveals image classification model pathologies. *Advances in Neural Information Processing Systems*, 34: 15395–15407.

Chormai, P.; Herrmann, J.; Müller, K.-R.; and Montavon, G. 2024. Disentangled explanations of neural network predictions by finding relevant subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Contier, O.; Baker, C. I.; and Hebart, M. N. 2024. Distributed representations of behaviour-derived object dimensions in the human visual system. *Nature Human Behaviour*, 8(11): 2179–2193.

Conwell, C.; Prince, J. S.; Kay, K. N.; Alvarez, G. A.; and Konkle, T. 2024. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1): 9383.

Dapello, J.; Marques, T.; Schrimpf, M.; Geiger, F.; Cox, D.; and DiCarlo, J. J. 2020. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33: 13073–13087.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feather, J.; Leclerc, G.; Mądry, A.; and McDermott, J. H. 2023. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11): 2017–2034.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, Y.; Shen, Z.; and Cui, P. 2021. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110: 107383.

Hebart, M. N.; Contier, O.; Teichmann, L.; Rockter, A. H.; Zheng, C. Y.; Kidder, A.; Corriveau, A.; Vaziri-Pashkam, M.; and Baker, C. I. 2023. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12: e82580.

Hermann, K.; Chen, T.; and Kornblith, S. 2020. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33: 19000–19015.

Hermann, K. L.; Mobahi, H.; Fel, T.; and Mozer, M. C. 2023. On the foundations of shortcut learning. *arXiv preprint arXiv:2310.16228*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Khosla, M.; Jamison, K.; Kuceyeski, A.; and Sabuncu, M. 2022. Characterizing the ventral visual stream with response-optimized neural encoding models. *Advances in Neural Information Processing Systems*, 35: 9389–9402.

Kriegeskorte, N.; Mur, M.; and Bandettini, P. A. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2: 249.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kubilius, J.; Schrimpf, M.; Kar, K.; Rajalingham, R.; Hong, H.; Majaj, N.; Issa, E.; Bashivan, P.; Prescott-Roy, J.; Schmidt, K.; et al. 2019. Brain-like object recognition with high-performing shallow recurrent ANNs. *Advances in neural information processing systems*, 32.

Li, Z.; Brendel, W.; Walker, E.; Cobos, E.; Muhammad, T.; Reimer, J.; Bethge, M.; Sinz, F.; Pitkow, Z.; and Tolias, A. 2019. Learning from brains how to regularize machines. *Advances in neural information processing systems*, 32.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Lopez-Cardona, A.; Segura, C.; Karatzoglou, A.; Abadal, S.; and Arapakis, I. 2024. Seeing Eye to AI: Human Alignment via Gaze-Based Response Rewards for Large Language Models. *arXiv preprint arXiv:2410.01532*.

Matsuyama, T.; Sasaki, K. S.; and Nishimoto, S. 2023. Applicability of scaling laws to vision encoding models. *arXiv preprint arXiv:2308.00678*.

Muttenthaler, L.; Greff, K.; Born, F.; Spitzer, B.; Kornblith, S.; Mozer, M. C.; MÃžller, K.-R.; Unterthiner, T.; and

Lampinen, A. K. 2024. Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*.

Muttenthaler, L.; Linhardt, L.; Dippel, J.; Vandermeulen, R. A.; Hermann, K.; Lampinen, A.; and Kornblith, S. 2023. Improving neural network representations using human similarity judgments. *Advances in Neural Information Processing Systems*, 36: 50978–51007.

Opiełka, G.; Loke, J.; and Scholte, S. 2024. Saliency Suppressed, Semantics Surfaced: Visual Transformations in Neural Networks and the Brain. *arXiv preprint arXiv:2404.18772*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Peer, M.; Brunec, I. K.; Newcombe, N. S.; and Epstein, R. A. 2021. Structuring knowledge with cognitive maps and cognitive graphs. *Trends in cognitive sciences*, 25(1): 37–54.

Ren, Y.; and Bashivan, P. 2024. How well do models of visual cortex generalize to out of distribution samples? *PLOS Computational Biology*, 20(5): e1011145.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

St-Yves, G.; Allen, E. J.; Wu, Y.; Kay, K.; and Naselaris, T. 2023. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature communications*, 14(1): 3329.

Steinmann, D.; Divo, F.; Kraus, M.; Wüst, A.; Struppek, L.; Friedrich, F.; and Kersting, K. 2024. Navigating Shortcuts, Spurious Correlations, and Confounders: From Origins via Detection to Mitigation. *arXiv preprint arXiv:2412.05152*.

Sucholutsky, I.; Muttenthaler, L.; Weller, A.; Peng, A.; Bobu, A.; Kim, B.; Love, B. C.; Grant, E.; Groen, I.; Achterberg, J.; et al. 2023. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Sundaram, S.; Fu, S.; Muttenthaler, L.; Tamir, N.; Chai, L.; Kornblith, S.; Darrell, T.; and Isola, P. 2024. When does perceptual alignment benefit vision representations? *Advances in Neural Information Processing Systems*, 37: 55314–55341.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tan, M. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; and Schmidt, L. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33: 18583–18599.

Wichmann, F. A.; and Geirhos, R. 2023. Are deep neural networks adequate behavioral models of human visual perception? *Annual review of vision science*, 9(1): 501–524.

Xu, Y.; and Vaziri-Pashkam, M. 2021. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications*, 12(1): 2065.

Yang, H.; Gee, J.; and Shi, J. 2024. Brain decodes deep nets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23030–23040.

Yang, Y.; Nushi, B.; Palangi, H.; and Mirzasoleiman, B. 2023. Mitigating spurious correlations in multi-modal models during fine-tuning. In *International Conference on Machine Learning*, 39365–39379. PMLR.

Zheng, J.; and Makar, M. 2022. Causally motivated multi-shortcut identification and removal. *Advances in Neural Information Processing Systems*, 35: 12800–12812.