
Mice to Machines: Neural Representations from Visual Cortex for Domain Generalization

Ahmed Qazi^{1*} Hamd Jalil^{1*} Asim Iqbal^{1†}

¹Tibbling Technologies
asim@tibbtech.com

Abstract

The mouse is one of the most studied animal models in the field of systems neuroscience. Understanding the generalized patterns and decoding the neural representations that are evoked by the diverse range of natural scene stimuli in the mouse visual cortex is one of the key quests in computational vision. In recent years, significant parallels have been drawn between the primate visual cortex and hierarchical deep neural networks. However, their generalized efficacy in understanding mouse vision has been limited. In this study, we investigate the functional alignment between the mouse visual cortex and deep learning models for object classification tasks. We first introduce a generalized representational learning strategy that uncovers a striking resemblance between the functional mapping of the mouse visual cortex and high-performing deep learning models on both top-down (population-level) and bottom-up (single cell-level) scenarios. Next, this representational similarity across the two systems is further enhanced by the addition of **Neural Response Normalization (NeuRN)** layer, inspired by the activation profile of excitatory and inhibitory neurons in the visual cortex. To test the performance effect of NeuRN on real-world tasks, we integrate it into deep learning models and observe significant improvements in their robustness against data shifts in domain generalization tasks. Our work proposes a novel framework for comparing the functional architecture of the mouse visual cortex with deep learning models. Our findings carry broad implications for the development of advanced AI models that draw inspiration from the mouse visual cortex, suggesting that AI models serve as valuable tools to explore neural representations of the visual cortex and, as a result, enhance their performance on real-world tasks.

1 Introduction

Deep learning has revolutionized machine learning, enabling breakthroughs across a range of applications, particularly in computer vision [Krizhevsky, Sutskever, and Hinton (2012); Simonyan and Zisserman (2015); He et al. (2016); Chollet (2017); Vaswani et al. (2017)]. Deep neural networks (DNNs) have achieved remarkable performance in tasks such as image recognition, object detection, and semantic segmentation by learning complex, high-dimensional representations from large datasets.

Despite these advances, a critical challenge that persists is the problem of *domain shift*, where models trained on a specific source domain perform poorly when applied to unseen target domains due to differences in data distributions [Blanchard et al. (2021); Taori et al. (2020); Ben-David et al. (2010); Recht et al. (2019); Moreno-Torres et al. (2012)]. This limitation hinders the deployment of DNNs in real-world scenarios, where data often vary widely and cannot be fully anticipated during training.

*Equal contribution.

†Corresponding author.

Addressing domain shift requires models that can generalize across diverse environments by capturing domain-agnostic features robust to variations in input data. Biological visual systems, particularly the mouse visual cortex (MVC), excel at this kind of generalization, efficiently processing a wide range of visual stimuli with remarkable adaptability. The MVC’s ability to extract invariant features from complex visual scenes offers valuable insights for improving domain generalization in artificial systems.

In this paper, we leverage principles from the MVC to tackle the domain shift problem in deep learning. Inspired from the Winner-Takes-All (WTA) mechanism observed in neuronal circuits (Iqbal et al. (2025)), we applied this approach to enable **NeuRN**—**Neural Response Normalization**—a biologically inspired technique designed to enhance both the domain generalization capabilities and biological alignment of DNNs. NeuRN normalizes pixel-level contrastive deviations to generate domain-agnostic feature representations, enabling generalization across varying domains. Furthermore, NeuRN increases the biological alignment between artificial neural networks and the MVC by incorporating properties of excitatory neurons into DNNs. Our contributions are:

- **Framework for Biological and Artificial Representation Comparison:** We introduce a framework for comparing neural representations from the MVC and feature representations from DNNs, revealing a strong alignment, particularly with excitatory neurons. This comparison is conducted at both the population level and the single-neuron level, providing a comprehensive understanding of the similarities and differences between biological and artificial systems. **Figure 1** shows the block diagram of our comparison framework.
- **Development of NeuRN:** Taking inspiration from the computational principles of visual neurons Iqbal et al. (2025), we applied it as a neuro-inspired normalization technique (NeuRN) that enhances domain generalization by capturing domain-agnostic features. By mimicking the response normalization property of excitatory and inhibitory neurons, NeuRN allows DNNs to focus on intrinsic structural features rather than superficial variations, improving their ability to generalize across different domains.
- **Evaluation of NeuRN’s Impact:** We evaluate NeuRN’s impact on both domain generalization tasks and biological alignment. Our experiments demonstrate that DNNs equipped with NeuRN not only perform better on challenging datasets such as MNIST, SVHN, USPS, and MNIST-M but also exhibit increased similarity to biological neural representations. We assess this alignment with and without NeuRN, showing that NeuRN enhances the biological plausibility of DNN representations.

By bridging the gap between biological and artificial neural systems, our work offers a novel approach to improving the robustness and adaptability of DNNs in the face of domain shift. Our findings have broad implications for the development of advanced AI models that draw inspiration from the visual processing capabilities of the brain, paving the way for more biologically plausible and generalizable artificial intelligence.

2 Related Work

Neuro-inspired models have been widely studied to solve the problem of domain generalization (Razzaq and Iqbal (2025b), Razzaq and Iqbal (2025a)). Convolutional Neural Networks (CNNs) have demonstrated the ability to predict neural responses in the MVC (Cadena et al. (2019b), Iqbal et al. (2019b)). Comparing DNNs with biological neural responses has provided valuable insights into how artificial systems can emulate biological visual systems Shi et al. (2022); Gao, Zou, and Duan (2023); Conwell et al. (2021). Furthermore, studies have shown that biasing CNN models toward a biological feature space results in more robust models Li et al. (2019). A significant body of work has focused on representational similarities between feature representations in DNNs and neural activity in response to various visual stimuli Yamins et al. (2014); Cadieu et al. (2014); Cadena et al. (2019a); Iqbal et al. (2019a); Shi, Shea-Brown, and Buice (2019); Conwell et al. (2021). Models such as CORnet Kubilius et al. (2018); Schrimpf et al. (2018) and frameworks like MouseNet Shi et al. (2022) and CVSNet Gao, Zou, and Duan (2023) have successfully incorporated biological principles, enabling DNNs to better mimic the efficiency and adaptability of the brain’s visual processing. For example, Li et al. (2019) leveraged DNNs to probe the MVC, utilizing innovative regularization techniques to increase robustness against adversarial attacks. While these studies have been pivotal in advancing our understanding of mammalian brain functionalities and inspiring DNNs, a common limitation is

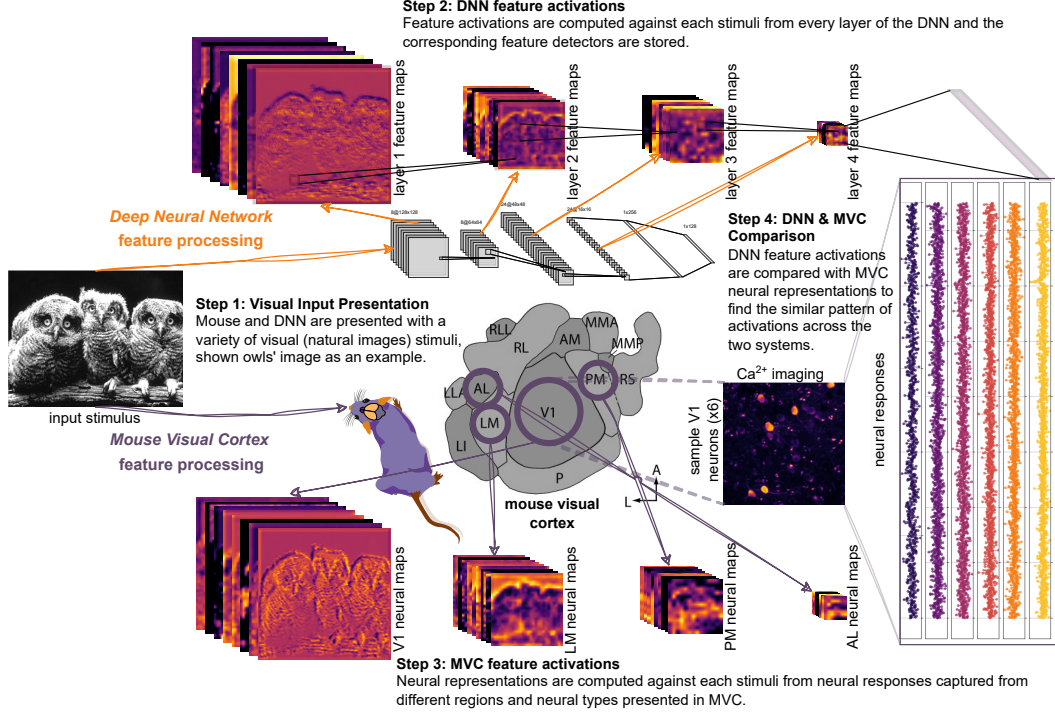


Figure 1: Block diagram architecture of comparison framework: image stimuli are presented to mice and DNN models (step 1) and their corresponding feature (step 2) and neural representations (step 3) are computed and systematically compared (step 4).

their scalability and applicability to diverse, real-world scenarios. Many proposed models, although theoretically profound, struggle to integrate effectively into practical applications across different domains. Building on this prior work, our framework compares MVC receptive fields (hereby referred to as neural representations) and DNN kernels (hereby referred to as feature representations), finding a strong similarity. This insight informs the design of NeuRN, which aligns DNN representations with biological plausibility, enabling robust generalization across varied visual tasks.

3 Methods

3.1 Allen Brain Observatory dataset

To conduct the comparison between DNNs and MVC, we utilized the Allen Brain Observatory open-source dataset [Allen Institute for Brain Science \(2016\)](#), which consists of in-vivo Calcium imaging data from GCaMP6-expressing neurons. These neurons were strategically selected from specific MVC areas, cortical layers, and Cre lines labeling a variety of neuron types. We specifically focused on the peak DF/F and time-to-peak characteristics. Furthermore, the mice in the experiment were presented with a variety of image stimuli, encompassing a wide spectrum of naturally occurring visual elements and environments (**Supplementary Figure S1**), to closely mimic the diversity of visual inputs they might encounter in the wild. Neural activity during this image exposure was recorded and collated, providing a rich dataset on the response patterns of mouse visual neurons to different visual stimuli. Using these neural traces, we constructed the neural representations dataset to serve as the foundational basis for our comparative analyses.

3.2 Neural representations dataset

Each stimulus from the Allen Brain Observatory natural scenes dataset [Allen Institute for Brain Science \(2016\)](#) was resized and flattened to create a uniform representation. This process resulted in a collection of flattened stimuli $FS = \{f_{s_1}, f_{s_2}, \dots, f_{s_n}\}$, where f_{s_i} is the flattened representation

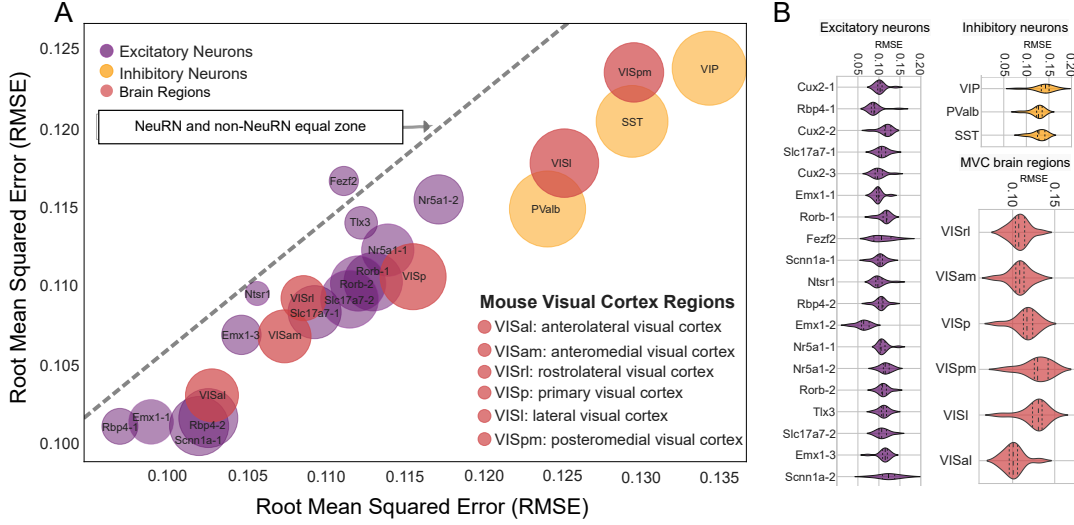


Figure 2: A) RMSE score comparison of feature representations from NeuRN and non-NeuRN DNN models with MVC neural representations. Data points below the diagonal line signify better NeuRN models’ biological plausibility with MVC neural representations. B) RMSE scores of feature representations’ comparisons with neural representations across neural types for 13 (non-NeuRN) DNN models.

of the i^{th} stimulus. For each neuron, we captured a series of trial traces in response to the stimulus presentations, forming $T = \{t_1, t_2, \dots, t_n\}$, where each t_i is a concatenated vector of trial traces corresponding to the i^{th} stimulus presentation. The neural representation N for each neuron was computed as the product of the transpose of the trial trace matrix T^T and the flattened stimulus matrix FS , resulting in:

$$N = T^T \cdot FS$$

Each element n_{ij} of the matrix N represents a 2D matrix that captures a detailed measure of the neuron’s response, including both intensity and spatial distribution.

We analyzed neural responses from 1,000 neurons (500 excitatory and 500 inhibitory) selected using Uniform Manifold Approximation and Projection (UMAP) embeddings [Hastie and McMahan \(2018\)](#) to capture the high-dimensional structure of the data. The UMAP parameters were set to $n_neighbors = 15$, prioritizing local structure preservation, with a ‘euclidean’ distance metric for similarity measurement, and $min_dist = 0.1$ to prevent excessive clustering. Following UMAP dimensionality reduction, we applied K-means clustering with $k = 10$ to segment the neurons into clusters that balance granularity with over-segmentation. From each cluster, we extracted 50 central neural representations for both excitatory and inhibitory neurons, ensuring representative samples while avoiding outliers. This approach allowed us to obtain a diverse yet representative subset of neurons for our analysis.

3.3 Artificial feature representations

To ensure that the DNNs have captured features from images similar to natural stimuli and are therefore comparable with neural representations from the MVC, we pre-trained the DNNs on ImageNet and fine-tuned them on a specifically curated subset of CIFAR-100 (please refer to the supplementary materials for further details on DNN model training and the CIFAR-100 subset). To extract artificial feature representations, we utilized the convolutional layer L of the DNNs, defined by dimensions $[F, C, H, W]$, where F signifies the number of filters, C the number of channels, H the height, and W the width. We considered each channel of every filter individually and reshaped it to $[F \times C, H, W]$, interpreting each channel of a filter as a distinct 2D feature representation.

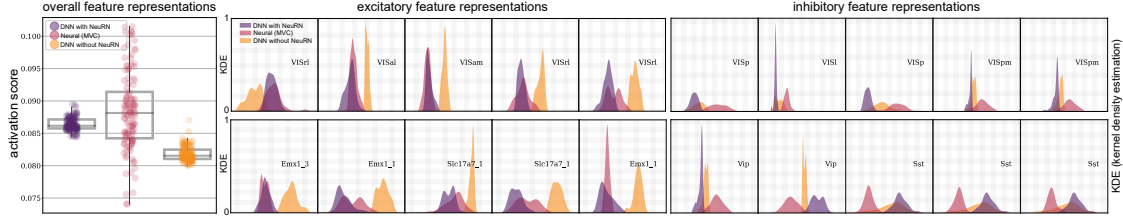


Figure 3: Distributional comparison of NeuRN and non-NeuRN feature representations with neural representations is shown. KDE curves of randomly sampled excitatory and inhibitory neurons show that NeuRN augmented feature representations are more similar to biological representations.

To select a representative sample of these feature representations for comparison with neural data, we applied UMAP embeddings to capture the underlying structure of the high-dimensional data. Subsequently, we performed K-means clustering (using the same parameters as for the biological neural representations) to group similar feature representations. This approach allowed us to generate a representative sample of 500 feature representations for each DNN, facilitating a meaningful comparison with the neural representations.

3.4 High-level representational analysis

To compare the neural and artificial feature representations, we considered an artificial feature representation K_1 and a biological neural representation K_2 , both of size $m \times n$. We reshaped K_1 and K_2 into 1D arrays:

$$\begin{aligned} K1_flat &= [k1_1, k1_2, \dots, k1_{mn}] \\ K2_flat &= [k2_1, k2_2, \dots, k2_{mn}] \end{aligned}$$

With $K1_flat$ and $K2_flat$ mapped to an mn -dimensional space, we computed the Root Mean Squared Error (RMSE) between these vectors as a measure of dissimilarity:

$$RMSE(K1_flat, K2_flat) = \sqrt{\frac{1}{mn} \sum_{i=1}^{mn} (k2_i - k1_i)^2} \quad (1)$$

In our comparison methodology, which is essentially Representational Similarity Analysis (RSA), we use RMSE as the distance metric. This choice aligns with approaches in previous studies, such as those by [Khaligh-Razavi and Kriegeskorte \(2014\)](#), where they discuss the use of Euclidean distance—closely related to RMSE—in RSA frameworks. We employ RMSE because it provides a direct and interpretable measure of the average magnitude of differences between corresponding elements in high-dimensional neural and model representations. RMSE captures both the direction and magnitude of discrepancies, making it sensitive to absolute differences in activation patterns. This sensitivity is crucial for our analysis, as we aim to assess not just the pattern similarity but also the fidelity of the model’s predicted neural responses to the actual biological neural data. By using RMSE, we can quantify how closely our models replicate the neural representations, accounting for both pattern shape and activation magnitude, thereby providing a robust metric for evaluating representational alignment.

3.5 Single neuron-level analysis

Traditional efforts, such as those in [Williams et al. \(2021\)](#); [Conwell et al. \(2021\)](#), have aimed to understand the representational similarity of neural manifolds by mapping them to the architectural space of models. These approaches primarily focus on embedding high-level representations of the neural population, often overlooking the intricacies at the single neuron level. Our work introduces a novel depth of exploration at this granular level, providing a richer understanding of the representational similarity between biological neural representations and artificial feature representations. In the MVC, the neural representations are used, captured by presenting specific stimuli. Analogously, in DNNs, feature representations act as the response to input stimuli. By comparing the feature representations

from both NeuRN and non-NeuRN DNNs, we can contrast every neural representation in our dataset with the feature representations from each DNN layer, ensuring a comprehensive comparison with minimal data loss.

Our methodology offers a detailed examination of neural compatibility, going beyond the generalized shape metrics on neural representations explored in works like Williams et al. (2021). While linear regression-based single neuron predictivity, as mentioned in Conwell et al. (2021), provides insights into neural representation correlations, our approach introduces a biophysical analysis that aligns feature representations at both the collective population and individual neuron levels. By emphasizing the distributional comparison of activations, we enrich traditional methodologies, extending beyond sole reliance on RSA. For each DNN layer, we retrieve its feature representations and evaluate their distributions using a Kernel Density Estimate (KDE). Similarly, we compute the KDE of the neural representations. The similarity between these KDEs is determined using the Intersection over Union (*IoU*) metric.

For each layer in a DNN, we extract the feature representations x of R dimensions and compute their distributions using a KDE:

$$K(u) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{u^2}{2\sigma^2}\right) \quad (2)$$

where u is the distance from the center of the kernel, and the factor

$$\frac{1}{\sqrt{2\pi}\sigma} \quad (3)$$

ensures that the kernel integrates to 1. The kernel density estimator for feature representations, $f(x)$, denoted as M :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (4)$$

To get the KDE for 2D neural representations (N) we repeat the same process. We calculate the overlap between biological and artificial KDEs to quantify their similarity, where *IoU* is derived as:

$$IoU = \frac{R \cap N}{R \cup N} \quad (5)$$

3.6 CIFAR-100 subset: CIFAR-27

Alongside our work with mice and natural scenes, we used models pre-trained on ImageNet, a large-scale and diverse image dataset widely used in deep learning. ImageNet pre-training enables the model to develop a broad set of feature detectors that can be applied across various visual tasks. We fine-tuned the models on a specifically curated subset of the CIFAR-100 dataset (**Supplementary Figure S2**). The CIFAR-100 dataset is a collection of 60,000 32x32 color images in 100 classes, with 600 images per class. Our subset comprised 27 classes, all pertaining to natural scenes or organisms. The classes included: butterfly, chimpanzee, bear, elephant, forest, fox, kangaroo, leopard, lion, lizard, maple_tree, mountain, oak_tree, orchid, otter, palm_tree, pine_tree, plain, poppy, rose, sea, sunflower, tiger, tulip, turtle, willow_tree, and wolf. This data set was developed to ensure proximity to natural scenes stimuli.

3.7 Evaluation datasets

Our domain generalization study utilizes diverse datasets including MNIST, SVHN, USPS, and MNIST-M, each presenting unique challenges LeCun, Cortes, and Burges (2010); Netzer et al. (2011); Hull (1994); Ganin et al. (2015). **MNIST** is a standard benchmark with 70,000 grayscale images of handwritten digits LeCun, Cortes, and Burges (2010). **SVHN**, with over 600,000 images, features house numbers in real-world scenarios, adding complexity to digit recognition tasks Netzer et al. (2011). **USPS**, comprising 9,298 16x16 grayscale images from scanned mail Hull (1994). Finally, **MNIST-M** fuses MNIST digits with BSDS500's colored backgrounds, testing models against variations in color and texture Ganin et al. (2015). These datasets effectively test machine learning models' ability to generalize from clean, handwritten digits to recognizing numbers in varied, naturalistic settings. They serve as critical benchmarks to assess the generalization capabilities of vision models in diverse visual environments.

4 Results & discussion

For structural clarity, we outline our analysis as follows. We first extract neural representations from the MVC and artificial feature representations from DNNs to evaluate representational similarity. Our analysis reveals that excitatory neurons exhibit a significantly stronger correlation with artificial feature representations compared to inhibitory neurons. Based on these findings, we designed NeuRN by drawing inspiration from the response normalization property of neurons. Integrating NeuRN with DNNs led to a further increase in the similarity between artificial and biological neural representations. To assess NeuRN’s efficacy in improving generalization, we conduct domain generalization experiments using datasets such as MNIST, SVHN, USPS, and MNIST-M [LeCun, Cortes, and Burges \(2010\)](#); [Netzer et al. \(2011\)](#); [Hull \(1994\)](#); [Ganin et al. \(2015\)](#), which present varying challenges in digit recognition tasks. Our results demonstrate that NeuRN not only brings artificial feature representations closer to biological ones but also enhances DNNs’ ability to capture domain-agnostic features, leading to improved generalization across different domains. The following sections detail each step of our analysis.

4.1 Strong alignment between feature and neural representations

Representational analysis was performed on neural representations and feature representations from the selected DNN models to quantify the similarity between biological and artificial responses to stimuli. This comparison established a benchmark for how closely the artificial models could mimic the biological neural representations. The analysis, shown in [Figure 2\(B\)](#), indicates that the biological neural representations and artificial feature representations have a significant degree of similarity, as evidenced by their RMSE scores. This result reaffirms the connection between biological and artificial neural networks, which has been previously suggested in studies such as [Kubilius et al. \(2018\)](#); [Shi et al. \(2022\)](#); [Kubilius et al. \(2019\)](#); [Jozwik et al. \(2019\)](#); [Kar et al. \(2019\)](#); [Conwell et al. \(2021\)](#). We measured the similarity across all brain regions in the MVC, as well as across different neural genotypes. Among all the brain regions studied, the anterolateral visual area (VISal) prominently stood out for its close resemblance to DNN feature representations. This result aligns with the findings of [Conwell et al. \(2021\)](#), which highlighted the distinct patterns in VISal that mirrored those captured by DNNs. Moreover, excitatory neurons consistently exhibited lower RMSE compared to inhibitory neurons ([Figure 2\(B\)](#)), emphasizing a stronger correlation between the activity of excitatory neurons and DNN feature representations.

These findings suggest that DNNs are more effective at capturing the representational characteristics of excitatory neurons in the MVC. This strong alignment supports the potential of leveraging biological insights to inform and improve artificial neural network designs.

4.2 Neural Response Normalization (NeuRN)

Building on our observation of a strong similarity between excitatory neurons and DNN feature representations, and after incorporating insights from Winner-Takes-All (WTA) mechanism observed in neuronal circuits [Iqbal et al. \(2025\)](#), we adapted Neural Response Normalization (NeuRN), a neuro-inspired layer that embeds the response normalization property of excitatory and inhibitory neurons into DNNs [Zayyad, Maunsell, and MacLean \(2023\)](#); [Burg et al. \(2021\)](#); [Sawada and Petrov \(2016\)](#); [Das and Ray \(2021\)](#). These neurons are distinguished by their ability to encode the structure and contrast of input images. By incorporating this biological functionality into DNNs through NeuRN, we aim to enhance the models’ ability to replicate the natural processing capabilities of the brain.

Given an input image denoted by $I \in \mathbb{R}^{W \times H \times C}$, where W , H , and C represent the width, height, and channels of the image, respectively, we aim to obtain a domain-agnostic representation I_a of this image. The procedure involves extracting patches of size k from each channel in I . Assuming a stride of 1, the total number of these patches equals the pixel count of the input. We represent the collection of these patches as $P = \{p_{k_1}, p_{k_2}, \dots, p_{k_n}\}$, where p_k is defined as the patch of size k surrounding the pixel x located at coordinates (i, j) within the image. We then compute the mean μ_{p_k} and standard deviation σ_{p_k} of each patch:

$$\mu_{p_k} = \frac{1}{k^2} \sum_{i,j \in p_k} x_{ij}, \quad \sigma_{p_k} = \sqrt{\frac{1}{k^2} \sum_{i,j \in p_k} (x_{ij} - \mu_{p_k})^2}$$

Finally, we normalize the standard deviation to obtain the domain-agnostic representation:

$$I_a = \frac{1}{c} \cdot \sigma_{p_k}, \quad \text{where } c = \max(\sigma)$$

By using the standard deviation, σ_{p_k} , NeuRN captures the contrast of local features, leading to representations that can generalize better across domains. Unlike Local Response Normalization (LRN) [Krizhevsky, Sutskever, and Hinton (2012)], which focuses on channel-wise activations using ratios of activations to neighbouring sums, and Local Contrast Normalization (LCN) [Jarrett et al. (2009)], which enhances individual pixel contrast but may miss the larger structural context, NeuRN considers spatial information of patches and maintains structural details throughout the image. This holistic, context-aware approach accommodates both local and global patterns, making NeuRN robust across varied visual settings. By aligning closely with neurobiological insights—specifically how neurons in the visual cortex encode contrasting features—NeuRN provides a more biologically plausible model for response normalization.

Table 1: The table compares the performance of image-classification DNNs across domain transfer tasks from the source to the target (source \rightarrow target) domains. The domains involved are MNIST (M), SVHN (S), USPS (U), and MNIST-M (MM). Entries in bold indicate an enhancement in performance due to the NeuRN adaptation, and underlined values are close-to-benchmark scores.

Models	M \rightarrow U	M \rightarrow S	M \rightarrow MM	U \rightarrow M	U \rightarrow S	U \rightarrow MM	S \rightarrow M	S \rightarrow U	S \rightarrow MM	MM \rightarrow M	MM \rightarrow U	MM \rightarrow S
VGG19+NeuRN	74.2	24.5	62.6	48.4	20.4	30.5	51.3	28.4	38.4	96.2	74.3	33.6
VGG19	70.9	13.7	39.7	66.2	13.9	19.9	47.2	32.5	33.9	94.3	76.7	22.5
EfficientNetB0+NeuRN	7.6	6.8	18.2	11.6	10.8	10.8	24.2	14.9	21.1	93.1	51.3	21.8
EfficientNetB0	8.6	7.4	18.4	9.4	8.0	10.6	34.6	12.0	18.6	96.8	17.8	12.5
DenseNet121+NeuRN	26.4	15.9	50.8	43.1	14.8	21.0	49.1	20.0	32.6	85.2	54.0	26.3
DenseNet121	74.3	11.7	38.4	39.5	11.4	20.9	34.1	22.1	31.5	96.6	73.3	21.3
ShuffleNet+NeuRN	7.0	56.1	71.1	22.6	36.2	27.2	32.1	27.5	28.9	84.8	7.6	50.4
ShuffleNet	9.3	19.0	14.1	44.9	12.2	13.0	26.1	31.1	19.7	38.7	3.2	18.3
Xception+NeuRN	61.9	20.9	59.7	41.3	15.9	25.8	48.8	16.4	37.7	95.7	31.8	33.1
Xception	62.4	19.4	56.0	20.9	11.9	21.1	42.7	14.2	34.2	97.9	77.8	25.6
NASNetMobile+NeuRN	45.4	21.1	47.5	25.5	10.4	16.9	47.6	16.5	35.1	95.6	58.3	31.1
NASNetMobile	33.0	12.7	36.0	24.8	10.1	21.4	46.0	23.8	32.6	90.0	63.7	20.3
ResNet50v2+NeuRN	31.4	23.1	59.6	28.4	14.1	20.7	50.5	18.8	35.7	96.4	55.7	31.3
ResNet50v2	78.7	21.6	57.8	11.1	8.6	10.5	48.0	15.6	33.7	97.2	78.1	17.2
Inceptionv3+NeuRN	52.6	22.6	62.2	50.8	14.4	31.3	47.9	19.7	37.6	96.1	75.5	34.0
Inceptionv3	69.8	16.4	57.8	34.2	8.9	24.9	53.5	28.0	39.0	97.2	82.0	15.0
MobileNetv2+NeuRN	20.9	16.3	45.0	34.2	10.6	17.5	48.9	15.0	36.2	96.2	88.5	33.9
MobileNetv2	8.8	16.1	45.9	32.6	9.3	17.8	45.9	17.7	30.8	94.2	44.9	20.8
ViT+NeuRN	61.9	16.2	44.7	36.6	14.1	25.3	33.2	25.5	26.5	87.8	58.9	23.5
ViT	69.4	13.1	32.8	26.2	9.4	15.6	45.2	36.5	30.6	95.5	64.4	21.8

4.3 NeuRN improves feature and neural representational alignment in DNNs

Our results show that after incorporating NeuRN into the models and subjecting them to the same training regimen (pretrained on ImageNet and fine-tuned on a subset of CIFAR-100), models with NeuRN achieved lower RMSE scores compared to those without NeuRN. This demonstrates NeuRN’s effectiveness in enhancing the models’ ability to mimic biological neural representations. As shown in **Figure 2(A)**, all genotypes, except for Fezf2, exhibited lower RMSE scores in NeuRN models compared to their non-NeuRN counterparts. This suggests that NeuRN’s normalization effect improves representational similarity and, consequently, biological plausibility among DNN architectures. Furthermore, we observe that feature representations of models correspond more closely with excitatory neural representations, characterized by genotypes such as Rbp4, Emx1, and Scnn1a, than with inhibitory neural representations, associated with genotypes like Sst and Vip. However, the differences are insignificant. Such findings align with existing biological literature, which emphasizes the pivotal role of these biological neurons in information processing and signal transmission within neural circuits [Markram et al. (2004)]. The adaptation of NeuRN further accentuates this correlation. These results underscore the potential of NeuRN in bridging the gap between artificial and biological neural representations, enhancing the biological plausibility of DNNs, and potentially leading to models that more accurately reflect neural processing in the brain.

4.4 NeuRN improves feature and neural representational alignment at single neuron-level

Our experiments demonstrate that NeuRN-activated feature representations exhibit a higher average activation mean compared to those without NeuRN. This increase in activation levels makes the model’s neurons more comparable to excitatory neurons in the biological brain, known for their significant role in signal transmission and higher firing rates. As shown in **Figure 3**, there is substantial overlap between the activation profiles (measured using KDE) of excitatory genotypes and brain regions and those of NeuRN feature representations, indicating a close alignment with the distribution of biological neural responses. This similarity reflects not just activation patterns but also key statistical properties like mean activation levels and variability, which are crucial for accurate information processing in neural networks.

These findings align with our high-level representational analysis, reinforcing NeuRN’s consistent impact across different analysis levels. By enhancing activation levels in DNNs to match those of excitatory neurons, NeuRN improves the models’ ability to emulate the functional characteristics of biological neural networks. Thus, incorporating NeuRN increases the biological plausibility of DNN feature representations, particularly among excitatory neurons. This enhanced alignment suggests that NeuRN enables DNNs to better capture the intrinsic properties of biological neurons, leading to models that are more robust in domain generalization tasks and more reflective of actual neural processing in the brain.

4.5 NeuRN-derived performance boost for domain generalization in DNNs

Using the digit datasets MNIST, SVHN, USPS, and MNIST-M, we test NeuRN’s domain-agnostic feature encoding capabilities and evaluate its impact on domain generalization. These datasets present unique challenges due to variations in image quality, style, and background noise, making them suitable for testing domain generalization capabilities. Our results, summarized in **Table 1**, show that DNNs fine-tuned on a single source domain and tested on various target domains exhibited significant improvement in cross-domain evaluation when equipped with NeuRN. This improvement is particularly evident in transitions from MNIST-M to SVHN, MNIST to SVHN, MNIST to MNIST-M, and USPS to SVHN, highlighting NeuRN’s strong domain bridging capability. The significance of these results lies in NeuRN’s ability to enhance the robustness of DNNs against domain shifts, a critical challenge in deploying machine learning models in real-world scenarios where data distributions often differ from training conditions. By normalizing input data in a biologically inspired manner, NeuRN enables models to focus on intrinsic structural features rather than superficial variations, leading to improved generalization across diverse datasets.

5 Conclusion

This work presents a comprehensive framework for comparing biological neural representations from the mouse visual cortex with artificial feature representations from deep neural networks, enabling analysis at both population and single-neuron levels across different brain regions and genotypes. Our framework reveals strong representational similarities between excitatory neurons in the MVC and DNN feature maps, providing new insights into the computational principles shared between biological and artificial visual systems.

Building on these findings, we implemented Neural Response Normalization (NeuRN), a biologically-inspired technique that captures the response normalization properties of visual cortex neurons. NeuRN enhances both the biological alignment of artificial networks and their domain generalization capabilities. Through extensive evaluation on multiple datasets including MNIST, SVHN, USPS, and MNIST-M, we demonstrate that NeuRN-equipped models achieve significant improvements in cross-domain performance while exhibiting closer correspondence to biological neural responses. The dual benefit of enhanced biological plausibility and improved practical performance positions NeuRN as a valuable contribution to neuro-inspired machine learning. Our approach successfully bridges neuroscience and artificial intelligence, demonstrating that biologically-motivated design principles can yield both scientific insights and practical advances. The framework’s modularity allows for application to diverse neural populations and extension to other sensory modalities.

Future work will test NeuRN on broader visual tasks like segmentation, detection, and view synthesis, advancing biologically inspired methods for more robust and interpretable AI.

References

- Allen Institute for Brain Science. 2016. Brain Observatory. https://observatory.brain-map.org/visualcoding/stimulus/natural_scenes.
- Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 101–108. ACM.
- Blanchard, G.; Deshmukh, A. A.; Dogan, U.; Lee, G.; and Scott, C. 2021. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(145): 1–41.
- Burg, M. F.; et al. 2021. Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, 17(6): e1009028.
- Cadena, S. A.; Denfield, G. H.; Walker, E. Y.; Gatys, L. A.; Tolia, A. S.; Bethge, M.; and Ecker, A. S. 2019a. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS computational biology*, 15(4): e1006897.
- Cadena, S. A.; Sinz, F. H.; Muhammad, T.; Froudarakis, E.; Cobos, E.; Walker, E. Y.; Reimer, J.; Bethge, M.; Tolia, A. S.; and Ecker, A. S. 2019b. How well do deep neural networks trained on object recognition characterize the mouse visual system.
- Cadiou, C. F.; Hong, H.; Yamins, D. L.; Pinto, N.; Ardila, D.; Solomon, E. A.; Majaj, N. J.; and DiCarlo, J. J. 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, 10(12): e1003963.
- Chollet, F. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800–1807.
- Conwell, C.; et al. 2021. Neural Regression, Representational Similarity, Model Zoology & Neural Taskonomy at Scale in Rodent Visual Cortex. In *NeurIPS Proceedings*.
- Das, A.; and Ray, S. 2021. Effect of Cross-Orientation Normalization on Different Neural Measures in Macaque Primary Visual Cortex. *Cerebral Cortex Communications*, 2(1): tgab009.
- Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2015. Domain-Adversarial Training of Neural Networks.
- Gao, R.; Zou, H.; and Duan, Z. 2023. CVSNet: A Computer Implementation for Central Visual System of The Brain. arXiv:2305.19492.
- Hastie, T.; and McMahan, B. J. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
- Hull, J. J. 1994. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5): 550–554.
- Iqbal, A.; Dong, P.; Kim, C.; and Jang, H. 2019a. Decoding Neural Responses in Mouse Visual Cortex through a Deep Neural Network. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE.
- Iqbal, A.; Dong, P.; Kim, C. M.; and Jang, H. 2019b. Decoding neural responses in mouse visual cortex through a deep neural network. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.
- Iqbal, A.; Mahmood, H.; Stuart, G. J.; Fishell, G.; and Honnuraiah, S. 2025. Biologically grounded neocortex computational primitives implemented on neuromorphic hardware improve vision transformer performance. *Proceedings of the National Academy of Sciences*, 122(41): e2504164122.
- Jarrett, K.; Kavukcuoglu, K.; Ranzato, M.; and LeCun, Y. 2009. What is the best multi-stage architecture for object recognition? *2009 IEEE 12th International Conference on Computer Vision*, 2146–2153.

- Jozwik, K.; et al. 2019. To find better neural network models of human vision, find better neural network models of primate vision. DiCarlo Lab.
- Kar, K.; et al. 2019. Evidence that recurrent circuits are critical to the ventral stream’s execution of core object recognition behavior. *Nature Neuroscience*, 22: 974–983.
- Khaligh-Razavi, S.-M.; and Kriegeskorte, N. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10(11): e1003915.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Neural Information Processing Systems*, volume 25.
- Kubilius, J.; Schrimpf, M.; Nayeibi, A.; Bear, D.; Yamins, D. L.; and DiCarlo, J. J. 2018. CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *bioRxiv*.
- Kubilius, J.; et al. 2019. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In *NeurIPS Proceedings*.
- LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST Handwritten Digit Database.
- Li, Z.; et al. 2019. Learning from brains how to regularize machines. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.
- Markram, H.; Toledo-Rodriguez, M.; Wang, Y.; Gupta, A.; Silberberg, G.; and Wu, C. 2004. Interneurons of the neocortical inhibitory system. *Nature Reviews Neuroscience*, 5(10): 793–807.
- Moreno-Torres, J. G.; Raeder, T.; Alaiz-Rodríguez, R.; Chawla, N. V.; and Herrera, F. 2012. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1): 327–336.
- Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.
- Razzaq, T.; and Iqbal, A. 2025a. Multimodal Neuromorphic Event-Frame Fusion in Domain-Generalized Vision Transformer for Dynamic Object Tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4684–4691.
- Razzaq, T.; and Iqbal, A. 2025b. NeuReg: Domain-Invariant Brain Registration Across Species and Modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5751–5761.
- Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 5401–5410.
- Sawada, T.; and Petrov, A. 2016. Emulating and predicting physiological results of neurons in the primary visual cortex (V1) based on the divisive normalization model. *Journal of Vision*, 16(12): 958.
- Schrimpf, M.; Kubilius, J.; Hong, H.; Majaj, N.; Rajalingham, R.; Issa, E. B.; Kar, K.; Bashivan, P.; Prescott-Roy, J.; Schmidt, K.; Yamins, D. L.; and DiCarlo, J. J. 2018. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*.
- Shi, J.; Shea-Brown, E.; and Buice, M. 2019. Comparison against task driven artificial neural networks reveals functional properties in mouse visual cortex. *Advances in Neural Information Processing Systems*, 32.
- Shi, J.; et al. 2022. MouseNet: A biologically constrained convolutional neural network model for the mouse visual cortex. *PLOS Computational Biology*.
- Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*, 1–14.
- Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; and Schmidt, L. 2020. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, volume 33, 10581–10591.

- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 6000–6010. Curran Associates Inc.
- Williams, A. H.; Kunz, E. M.; Kornblith, S.; and Linderman, S. W. 2021. Generalized Shape Metrics on Neural Representations.
- Yamins, D. L.; Hong, H.; Cadieu, C. F.; Solomon, E. A.; Seibert, D.; and DiCarlo, J. J. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23): 8619–8624.
- Zayyad, Z. A.; Maunsell, J. H. R.; and MacLean, J. N. 2023. Normalization in mouse primary visual cortex. *bioRxiv*. The complete version history of this preprint is available at bioRxiv.