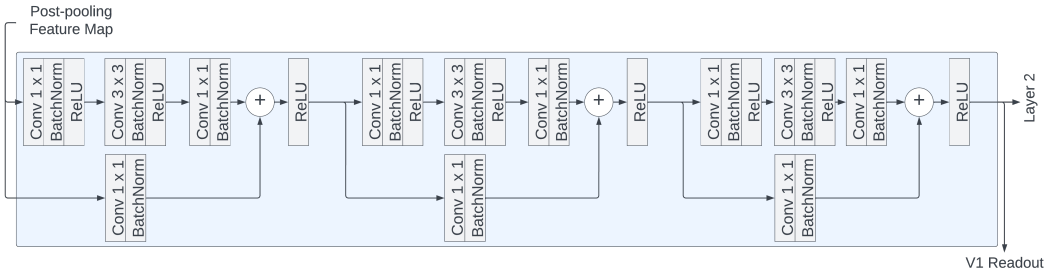
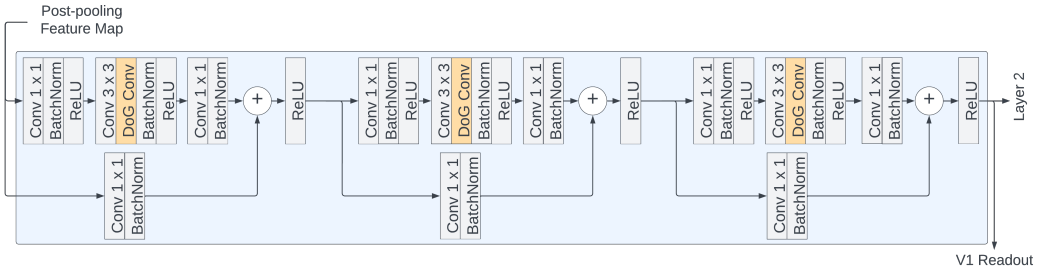
Appendix

A Supplemental Model Diagrams

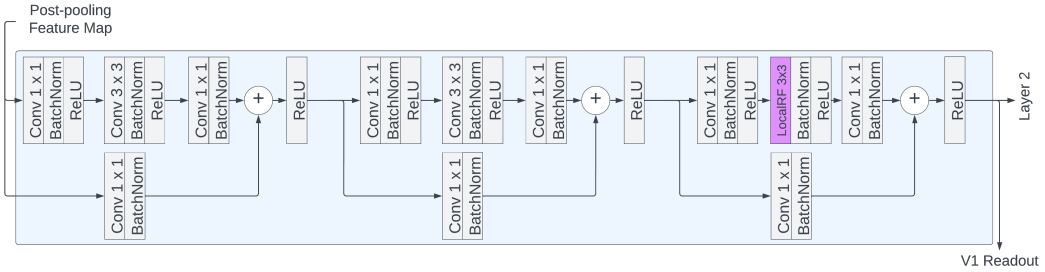
Fig. A.1 depicts the modifications made to ResNet50 residual layer 1 in the isolated component analyses of section 4.1. All multi-component (composite) models analyzed in Section 4.2 relied on combinations of these modifications (as exemplified in Fig. 2).



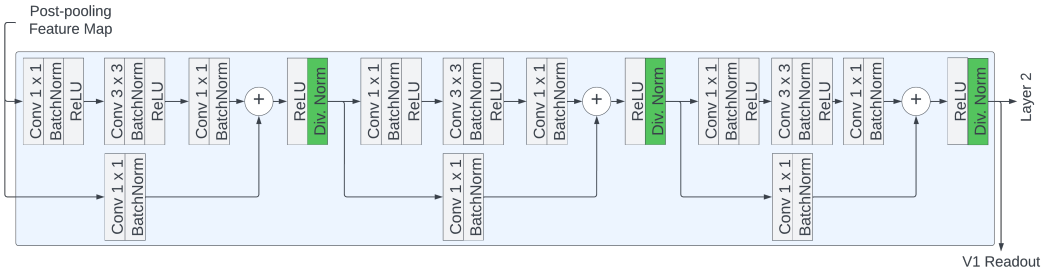
(A) ResNet50 baseline



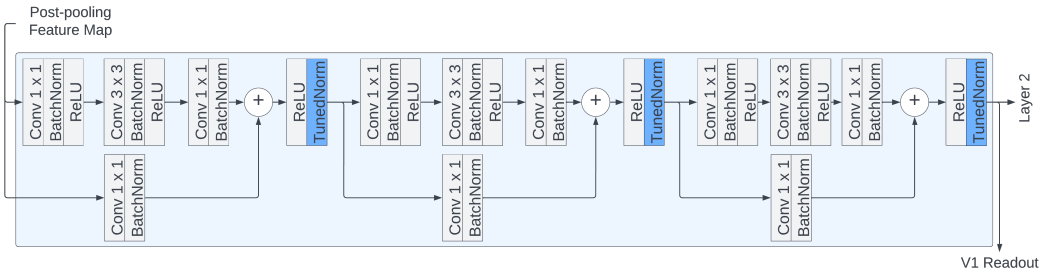
(B) Center-surround antagonism



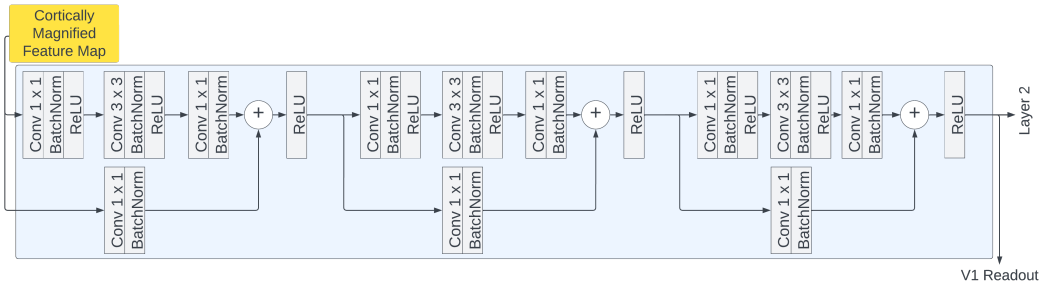
(C) Local receptive fields



(D) Divisive normalization



(E) Tuned normalization



(F) Cortical magnification

Figure A.1: ResNet50 residual layer 1, supplemented with individual neuro-constrained architectural components, as in section 4.1. (A) No modification (baseline ResNet50 layer 1), (B) with center-surround antagonism, (C) with local receptive field (RF), (D) with divisive normalization, (E) with tuned normalization, (F) with cortical magnification.

B V1 Scores of Alternate Layers of Baseline Network

When evaluating a model on Brain-Score, users are permitted to commit a mapping between model layers and areas of the ventral stream. Model-brain alignment is computed for each mapped pair in the Brain-Score evaluation. To promote a fair evaluation, we sought to find the layer that yielded optimal V1 alignment from the baseline ResNet50 model and fix this layer as the artificial V1 readout layer in all of our tested models. It is worth noting that after supplementing the base ResNet50 with neuro-constrained components, this layer may no longer offer optimal V1 alignment in the augmented network. In spite of this, we maintain this layer as our artificial V1 readout layer for fair evaluation.

To find the ResNet50 layer with the best V1 Overall, Predictivity, and Property scores, we compared a total of 20 different hidden layers (Fig. B.1). 16 of these layers corresponded to the post-activation hidden states of the network. The remaining 4 were downsampling layers of the first bottleneck block of each residual layer in the network, as these have previously demonstrated good V1 alignment [25]. Aside from these downsampling layers, hidden layers that did not follow a ReLU activation were omitted from this evaluation as the activities of these states can take on negative values and are therefore less interpretable as neural activities. Among all evaluated layers, the final output of ResNet50 residual layer 1 (i.e., the output of the third residual block of ResNet50) offered the highest V1 Overall score, and was therefore selected as the artificial V1 readout layer in all of our experiments.

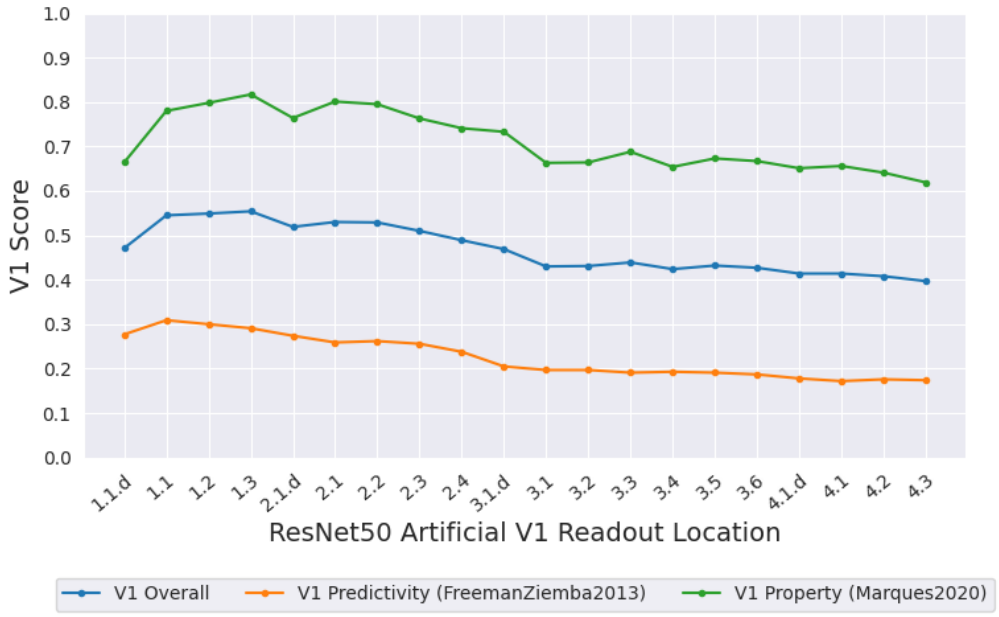


Figure B.1: V1 alignment Brain-Scores for 20 different hidden layers of ResNet50. In the plot above, readout location 'X.Y' denotes that artificial V1 activity was evaluated from residual block 'Y' of residual layer 'X'. Readout location suffixed with '.d' correspond to downsampling layers of the associated residual bottleneck. Highest V1 overall score came from block 3 of residual layer 1.

C Expanded Model Tuning Properties

Primary visual cortex (V1) tuning property alignments for each composite model evaluated in Section 4.2 are presented in Table 5. Tuning property similarities are computed as ceiled Kolmogorov-Smirnov distance between artificial neural response distributions from the model and empirical distributions recorded in primates [16, 17].

Center-Surround	Local RF	Tuned Norm.	Cortical Mag.	Adv. Training	Orientation	Spatial Frequency	Response Selectivity	RF Size	Surround Modulation	Texture Modulation	Response Mag.
✓	✓	✓	✓		.891	.925	.756	.840	.779	.844	.930
	✓	✓	✓		.858	.919	.780	.834	.808	.871	.930
✓		✓	✓		.894	.932	.750	.851	.775	.858	.946
✓	✓		✓		.878	.873	.739	.816	.719	.802	.910
✓	✓	✓			.875	.873	.702	.808	.890	.815	.870
		✓	✓		.873	.886	.735	.840	.794	.825	.959
	✓		✓		.902	.866	.715	.801	.625	.841	.869
	✓	✓			.915	.817	.691	.811	.898	.802	.911
✓	✓	✓	✓	✓	.924	.863	.773	.797	.733	.815	.899
	✓	✓	✓	✓	.944	.834	.768	.806	.673	.811	.900

Table 5: Ablation study model alignment across the seven primary visual cortex (V1) tuning properties that constitute the V1 Property score ('Marques2020') of Brain-Score. Checkmarks denote whether or not the architectural component was included in the model.

D V1 Brain-Scores of Untrained Models

	V1 Overall	V1 Predictivity	V1 Property
Center-surround antagonism	.298	.245	.551
Local receptive fields	.477	.210	.743
Divisive normalization	.499	.207	.792
Tuned normalization	.471	.218	.724
Cortical magnification	.497	.276	.718
All Components	.483	.225	.741
ResNet50	.466	.223	.710

Table 6: primary visual cortex (V1) alignment scores of untrained ResNet50 model variants.

E V2, V4, and IT Brain-Scores of Top Model

Table 7 shows the Brain-Scores of our top performing V1 model (the adversarially trained ResNet50 with all architectural components) for brain areas V2, V4, and IT. Network layers were mapped to visual areas V2, V4, and IT by finding the layers that achieve the best scores on these visual area benchmarks, as evaluated on Brain-Score's publicly available evaluation set.

Visual Area Brain-Score	V2	V4	IT
	.298	.245	.551

Table 7: V2, V4, and IT Brain-Scores of adversarially trained ResNet50 with all architectural components.

F Supplemental Visualizations

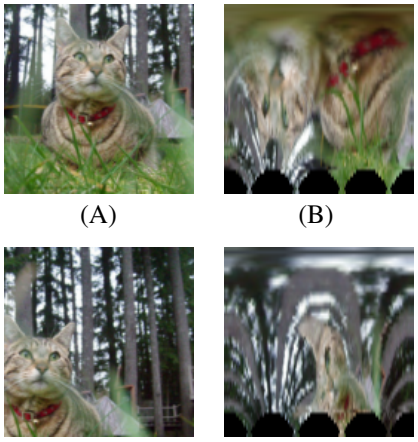


Figure F.1: Cortical magnification disproportionately samples the input image (or feature map). (A) Original image in which the object of interest (cat) is centered in the image frame. (B) Image (A) after the simulated cortical magnification transform. (C) Different crop of the original image, with the cat offset from image center. (D) Image (C) after the simulated cortical magnification transform

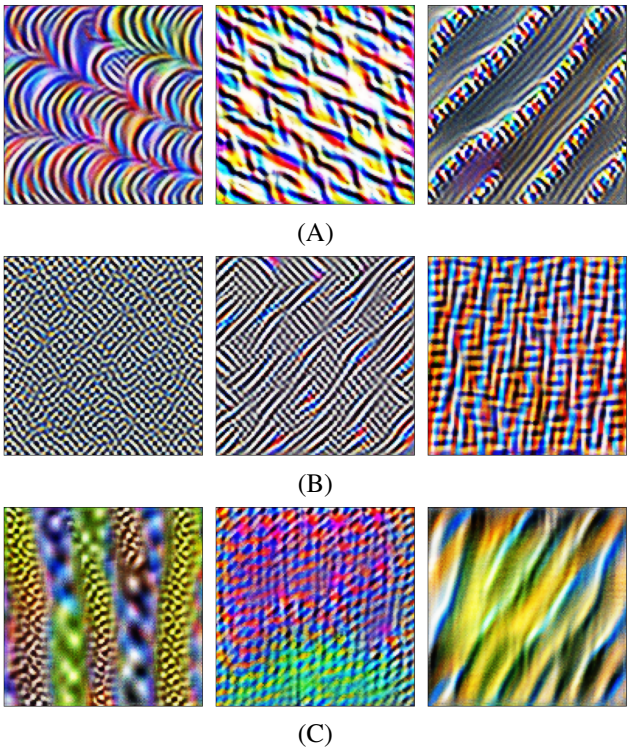


Figure F.2: Visual stimuli that maximally activate artificial V1 neurons of (A) baseline, (B) center-surround, and (C) tuned normalized ResNet50 networks. Maximally activating images generated using the python package 'lucent' (`https://https://github.com/greentfrapp/lucent`).

G Adversarial Training

The neuro-constrained ResNets discussed in Section 4.5 were trained using the "Free" adversarial training method proposed by Shafahi *et al.* [61]. In Projected Gradient Descent (PGD)-based adversarial training (a typical approach to adversarially training robust classifiers), a network is trained on adversarial samples that are generated on the fly during training. Specifically, in PGD-based adversarial training, a batch of adversarial images is first generated through a series of iterative perturbations to an original image batch, at which point the parameters of the network are finally updated according to the network's loss, as evaluated on the adversarial examples. "Free" adversarial training generates adversarial training images with a similar approach, but the parameters of the network are simultaneously updated with every iteration of image perturbation, significantly reducing training time. The authors refer to these mini-batch updates as "replays", and refer to the number of replays of each mini-batch with the parameter m.

The adversarially trained models of Section 4.5 were trained with $m = 4$ replays and perturbation clipping of $\epsilon = \frac{2}{255}$. These models were trained for 120 epochs using a stochastic gradient descent optimizer with an initial learning rate of 0.1, which was reduced by a factor of 10 every 40 epochs, momentum of 0.9, and weight decay of 1×10^{-5}. Each model was initialized with the weights that were learned during traditional ImageNet training for the analyses in Section 4.2. "Free" adversarial training was performed using code provided by the authors of this method (`https://github.com/mahyarnajibi/FreeAdversarialTraining`).

H Robustness to Common Image Corruptions

H.1 Dataset Description

We evaluated image classification robustness to common image corruptions using the Tiny-ImageNet-C dataset [56]. Recall that Tiny-ImageNet-C was used instead of ImageNet-C, because our models were trained on 64×64 input images. Downscaling ImageNet-C images would have potentially altered the intended corruptions and biased our evaluations.

Tiny-ImageNet-C is among a collection of corrupted datasets (e.g., ImageNet-C, CIFAR-10-C, CIFAR-100-C) that feature a diverse set of corruptions to typical benchmark datasets. Hendrycks and Dietterich [56] suggest that given the diversity of corruptions featured in these datasets, performance on these datasets can be seen as a general indicator of model robustness. The Tiny-ImageNet-C evaluation dataset consists of images from that Tiny-ImageNet validation dataset that have been corrupted according to 15 types of image corruption, each of which is categorized as a 'noise', 'blur', 'weather', or 'digital' corruption. The 15 corruption types include: Gaussian noise, shot noise, impulse noise, defocus blur, frosted glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic transformation, pixelation, and JPEG compression. Each corruption is depicted in Fig. H.1. Every image of this evaluation dataset is also corrupted at five levels of severity (the higher the corruption severity, the more the original image had been corrupted). Corruption severities for Gaussian noise are exemplified in Fig. H.2.

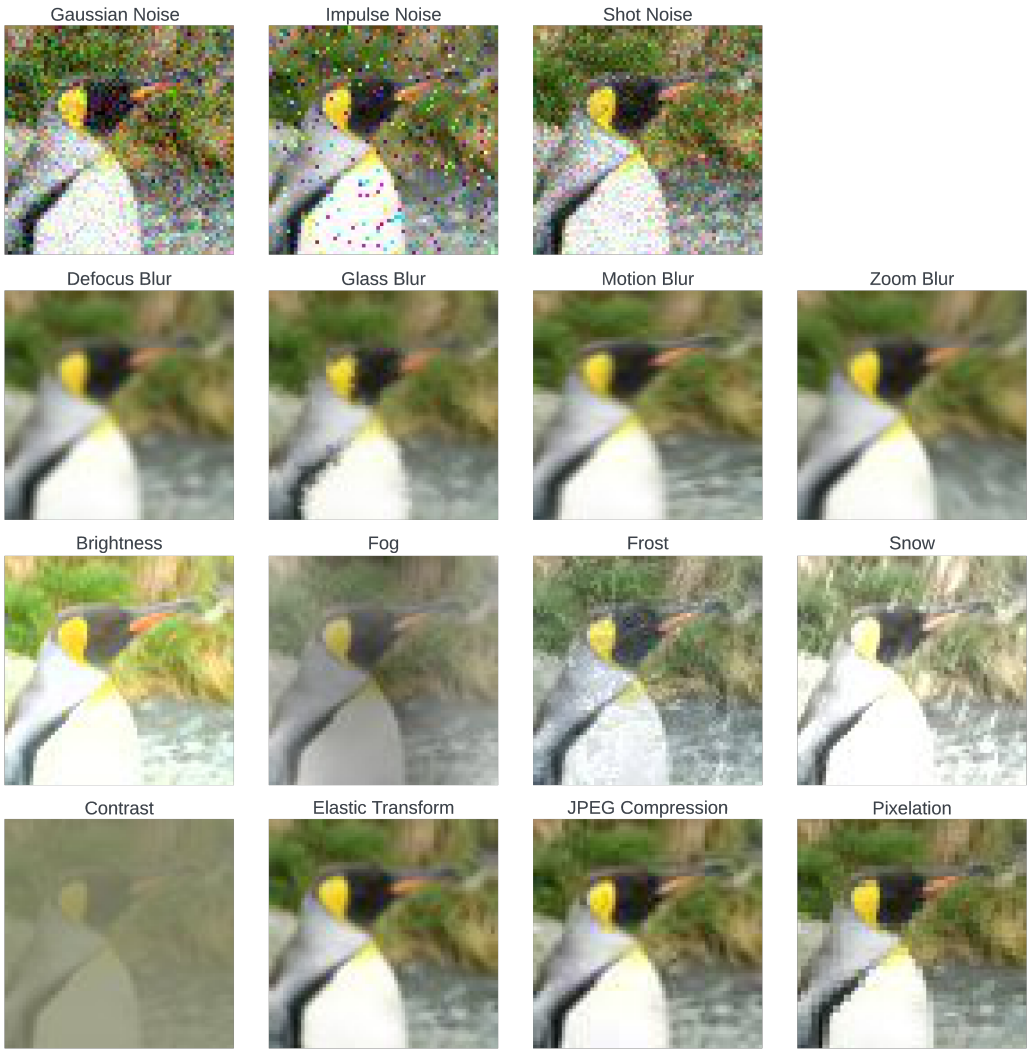


Figure H.1: 15 corruptions of the Tiny-ImageNet-C dataset, applied to a sample image from Tiny-ImageNet-C. First row: noise corruptions, second row: blur corruptions, third row: weather corruptions, bottom row: digital corruptions. All corruptions shown at severity level 3.

Figure H.2: Gaussian noise corruption, shown at corruption severity levels 1-5.

H.2 Corrupted Image Robustness

A detailed breakdown of Tiny-ImageNet-C image classification accuracy for each single-component, neuro-constrained ResNet-50 and the composite models that achieved top V1 Overall score without adversarial training are provided in Tables 8, 9, and 10.

	Tiny-ImageNet Val.	Tiny-ImageNet-C	Δ
ResNet50 (Baseline)	$\mathbf{.742} \pm .003$	$.278 \pm .004$	$.463 \pm .006$
Center-surround antagonism	$.739 \pm .004$	$.277 \pm .008$	$.463 \pm .009$
Local Receptive Fields	$.741 \pm .002$	$.275 \pm .004$	$.467 \pm .004$
Tuned Normalization	$.740 \pm .001$	$\mathbf{.283} \pm .006$	$\mathbf{.457} \pm .006$
Cortical Magnification	$.683 \pm .001$	$.222 \pm .009$	$.461 \pm .009$
Composite Model A	.694	.231	.463
Composite Model B	.691	.232	.459

Table 8: Classification accuracy of models on Tiny-ImageNet validation and Tiny-ImageNet-C (all corruption types and severities) datasets. Composite Model A includes all 4 neuro-constrained architectural components (center-surround antagonism, local receptive fields, tuned normalization, and cortical magnification). Composite Model B contained all architectural components, with the exception of center-surround antagonism. For baseline and single-component models, mean accuracies ($\pm$ one standard deviation) are reported, where each trial was associated with a distinct base model from the repeated trials of section 4.1.

	Corruption Severity				
	1	2	3	4	5
ResNet50 (Baseline)	$.418 \pm .004$	$.345 \pm .005$	$.269 \pm .004$	$.204 \pm .005$	$.156 \pm .003$
Center-surround antagonism	$.414 \pm .010$	$.343 \pm .009$	$.267 \pm .009$	$.203 \pm .006$	$.156 \pm .004$
Local Receptive Fields	$.416 \pm .003$	$.341 \pm .003$	$.264 \pm .003$	$.199 \pm .002$	$.153 \pm .002$
Tuned Normalization	$\mathbf{.424} \pm .006$	$\mathbf{.350} \pm .006$	$\mathbf{.274} \pm .007$	$\mathbf{.208} \pm .006$	$\mathbf{.160} \pm .004$
Cortical Magnification	$.349 \pm .011$	$.277 \pm .013$	$.208 \pm .010$	$.157 \pm .007$	$.120 \pm .006$
Composite Model A	.363	.289	.216	.163	.125
Composite Model B	.361	.288	.219	.165	.127

Table 9: Classification accuracy of models on Tiny-ImageNet-C at each level of corruption severity. Composite Model A includes all 4 neuro-constrained architectural components (center-surround antagonism, local receptive fields, tuned normalization, and cortical magnification). Composite Model B contained all architectural components, with the exception of center-surround antagonism. For baseline and single-component models, mean accuracies ($\pm$ one standard deviation) are reported, where each trial was associated with a distinct base model from the repeated trials of section 4.1.

Noise Corruptions

	Gaussian Noise	Impulse Noise	Shot Noise	Avg.
ResNet50 (Baseline)	$\mathbf{.197} \pm .011$	$.191 \pm .010$	$\mathbf{.232} \pm .013$	$\mathbf{.207} \pm .011$
Center-surround antagonism	$.195 \pm .010$	$.186 \pm .009$	$\mathbf{.232} \pm .012$	$.204 \pm .010$
Local Receptive Fields	$.185 \pm .006$	$.184 \pm 009$	$.219 \pm .010$	$.196 \pm .008$
Tuned Normalization	$.195 \pm .008$	$\mathbf{.192} \pm .004$	$.228 \pm .007$	$.205 \pm .006$
Cortical Magnification	$.150 \pm .008$	$.157 \pm .007$	$.180 \pm .011$	$.162 \pm .008$
Composite Model A	.151	.156	.184	.164
Composite Model B	.144	.149	.177	.157

Blur Corruptions

	Defocus Blur	Glass Blur	Motion Blur	Zoom Blur	Avg.
ResNet50 (Baseline)	$.224 \pm .003$	$.182 \pm .001$	$.272 \pm .003$	$.241 \pm .004$	$.230 \pm .002$
Center-surround antagonism	$.223 \pm .009$	$.184 \pm .004$	$.274 \pm .012$	$.243 \pm .011$	$.231 \pm .009$
Local Receptive Fields	$.228 \pm .006$	$.183 \pm .004$	$.273 \pm .005$	$.243 \pm .008$	$.232 \pm .005$
Tuned Normalization	$\mathbf{.234} \pm .009$	$\mathbf{.188} \pm .002$	$\mathbf{.277} \pm .009$	$\mathbf{.248} \pm .010$	$\mathbf{.237} \pm .007$
Cortical Magnification	$.174 \pm .010$	$.162 \pm .008$	$.222 \pm .007$	$.190 \pm .006$	$.187 \pm .008$
Composite Model A	.186	.167	.236	.200	.197
Composite Model B	.196	.174	.249	.222	.210

Weather Corruptions

	Brightness	Fog	Frost	Snow	Avg.
ResNet50 (Baseline)	$.401 \pm .005$	$\mathbf{.282} \pm .003$	$.360 \pm .006$	$.310 \pm .004$	$.338 \pm .004$
Center-surround antagonism	$.399 \pm .008$	$.270 \pm .008$	$.357 \pm .012$	$.302 \pm .003$	$.332 \pm .007$
Local Receptive Fields	$.398 \pm .008$	$.275 \pm .005$	$.351 \pm .006$	$.298 \pm .004$	$.331 \pm .003$
Tuned Normalization	$\mathbf{.410} \pm .008$	$\mathbf{.282} \pm .011$	$\mathbf{.361} \pm .006$	$\mathbf{.311} \pm .010$	$\mathbf{.341} \pm .008$
Cortical Magnification	$.327 \pm .011$	$.211 \pm .013$	$.283 \pm .014$	$.248 \pm .010$	$.267 \pm .011$
Composite Model A	.338	.220	.286	.258	.275
Composite Model B	.327	.225	.284	.255	.273

Digital Corruptions

	Contrast	Elastic	JPEG	Pixelate	Avg.
ResNet50 (Baseline)	$.125 \pm .001$	$.331 \pm .007$	$.454 \pm .007$	$.374 \pm .003$	$.321 \pm .003$
Center-surround antagonism	$.122 \pm .002$	$.331 \pm .014$	$.455 \pm .007$	$.374 \pm .004$	$.321 \pm .006$
Local Receptive Fields	$.120 \pm .004$	$.329 \pm .003$	$.457 \pm .005$	$.375 \pm .002$	$.320 \pm .001$
Tuned Normalization	$\mathbf{.128} \pm .008$	$\mathbf{.342} \pm .010$	$\mathbf{.463} \pm .006$	$\mathbf{.387} \pm .006$	$\mathbf{.330} \pm .007$
Cortical Magnification	$.082 \pm .005$	$.287 \pm .007$	$.374 \pm .013$	$.287 \pm .014$	$.257 \pm .010$
Composite Model A	.081	.305	.397	.303	.272
Composite Model B	.086	.314	.383	.293	.269

Table 10: Corrupted image classification accuracy by corruption type. Composite Model A includes all 4 neuro-constrained architectural components (center-surround antagonism, local receptive fields, tuned normalization, and cortical magnification). Composite Model B contained all architectural components, with the exception of center-surround antagonism. For baseline and single-component models, mean accuracies ($\pm$ one standard deviation) are reported, where each trial was associated with a distinct base model from the repeated trials of section 4.1.

I Code Availability

Code and materials required to reproduce the work presented in this paper are available at `github.com/bionicvisionlab/2023-Pogoncheff-Explaining-V1-Properties`.