SELF-ATTENTION-BASED CONTEXTUAL MODULATION IMPROVES NEURAL SYSTEM IDENTIFICATION

Isaac Lin[1,*], **Tianye Wang**[2], **Shang Gao**[1,3], **Shiming Tang**[2], **Tai Sing Lee**[1,*]
[1]Carnegie Mellon University, [2]Peking University, [3]Massachusetts Institute of Technology

ABSTRACT

Convolutional neural networks (CNNs) have been shown to be state-of-the-art models for visual cortical neurons. Cortical neurons in the primary visual cortex are sensitive to contextual information mediated by extensive horizontal and feedback connections. Standard CNNs integrate global contextual information to model contextual modulation via two mechanisms: successive convolutions and a fully connected readout layer. In this paper, we find that self-attention (SA), an implementation of non-local network mechanisms, can improve neural response predictions over parameter-matched CNNs in two key metrics: tuning curve correlation and peak tuning. We introduce peak tuning as a metric to evaluate a model's ability to capture a neuron's top feature preference. We factorize networks to assess each context mechanism, revealing that information in the local receptive field is most important for modeling overall tuning, but surround information is critically necessary for characterizing the tuning peak. We find that self-attention can replace posterior spatial-integration convolutions when learned incrementally, and is further enhanced in the presence of a fully connected readout layer, suggesting that the two context mechanisms are complementary. Finally, we find that decomposing receptive field learning and contextual modulation learning in an incremental manner may be an effective and robust mechanism for learning surround-center interactions.

1 INTRODUCTION

Feedforward CNN models have been shown in recent years to be an effective approach for modeling and predicting visual cortical neurons' responses to arbitrary natural images (Kindel et al., 2019; Klindt et al., 2018; Yamins & DiCarlo, 2016; Zhang et al., 2018; Cadena et al., 2017; Kriegeskorte, 2015). Neurons in the primate visual cortex are known to have extensive horizontal and feedback recurrent connections for mediating contextual modulation (Felleman & van Essen, 1991; Markov et al., 2014). Feedforward CNNs can model the influence of contextual surround on the responses of the neurons via two mechanisms: successive convolution layers and a fully connected layer. Both can make the neural model's responses sensitive to the global image context, outside the traditional classical receptive fields of neurons. Interestingly, feedforward CNNs trained on image classification tasks have also been shown to exhibit contextual surround modulation similar to what has been observed in neurophysiological experiments of the visual cortex (Pan et al., 2023).

In the context of neural prediction, it is found that including the inductive bias of horizontal recurrent connections can improve the model's predictive capabilities (Zhang et al., 2022). On the other hand, replacing a feedforward layer with a recurrent layer using a Markovian local kernel consistently outperforms parameter-matched feedforward CNNs in image classification tasks (Han et al., 2018; Nayebi et al., 2018; Kubilius et al., 2019; Zhang et al., 2022). However, contextual modulation in the visual cortex involves both the near surround and far surround, with the far surround being mediated by top-down feedback (Angelucci & Bressloff, 2006; Sasaki et al., 2013; Shushruth et al., 2013). In addition, there is evidence that contextual modulation is dynamic and highly image-dependent, suggesting a flexible gating mechanism (Coen-Cagli et al., 2015). Such a flexible gating mechanism can be modeled by a combination of Gaussian mixture models, implemented either by image-dependent normalization (Coen-Cagli et al., 2015) or by non-local networks and the self-attention mechanism

[*]Corresponding authors: Isaac Lin (isaacl@cs.cmu.edu), Tai Sing Lee (taislee@cmu.edu).

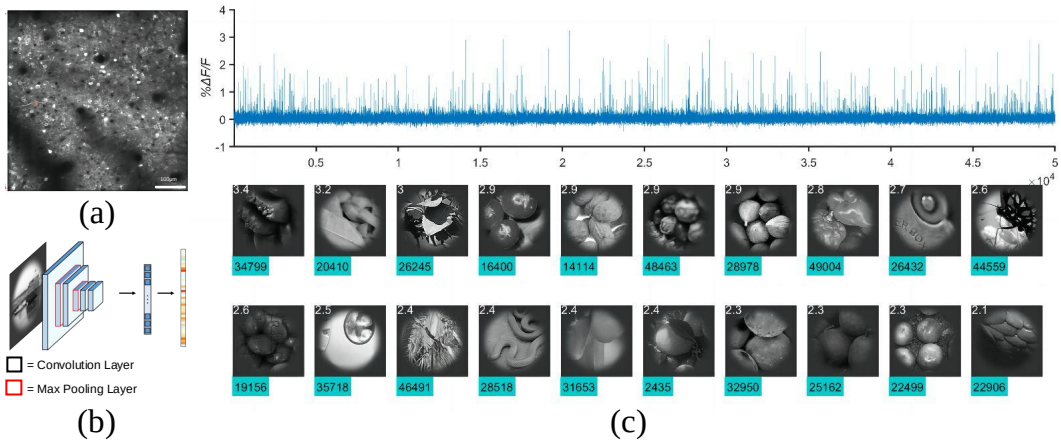


Figure 1: Macaque neuronal response dataset. **(a)** shows a two-photon image with cells. **(b)** shows a feedforward CNN used to model neural response. **(c)** shows the response of one neuron to 50k stimuli and the top 20 images that induced the strongest responses. On average, less than 0.5% of the images induce responses greater than half peak height. Each site contains around 300 neurons.

in deep learning (Fei & Pitkow, 2022). Self-attention-based architectures such as vision transformers have recently been shown to be effective in modeling mouse V1 neurons (Li et al., 2023). However, these networks are complex, with a large number of layers and multiple attention heads, making it difficult to isolate the key factors contributing to their superior performance.

In this paper, we demonstrate that adding a simple self-attention layer to a CNN can improve neural response prediction of macaque V1 neurons in two performance metrics: overall tuning correlation and prediction of the tuning peaks. To understand the mechanism driving improvement, we assessed the three contextual modulation mechanisms – convolutions, self-attention, and a fully connected readout layer. We found that while the three context mechanisms complement one another to produce the best prediction performance when used in conjunction, they have specific roles. First, the fully connected layer plays a critical role in peak prediction, though self-attention can further enhance it. Second, self-attention alone can improve tuning curve correlation but is insufficient for predicting the response peak or top feature preference. Third, the performance of self-attention models can be greatly enhanced when the feedforward receptive fields are learned first before learning the self-attention network, rather than learning everything from scratch simultaneously. The benefits of such incremental learning (van de Ven et al., 2022; Cotton et al., 2020) in this context suggest that decoupling the learning of feedforward receptive fields and recurrent connections may allow the system to learn a richer representation of contextual modulation. These observations provide new insights into the underlying computational mechanisms of cortical development and organization.

2 Related Works

Modeling neural response prediction Feedforward deep neural networks have proven effective in modeling and predicting neural responses in early visual brain areas (Kindel et al., 2019; Klindt et al., 2018; Yamins & DiCarlo, 2016; Zhang et al., 2018; Cadena et al., 2017; Kriegeskorte, 2015). However, the brain's visual areas contain abundant recurrent connections that are essential for generating neural responses (Felleman & van Essen, 1991; Markov et al., 2014; Spoerer et al., 2020). Incorporating biologically-inspired simple recurrent circuits, in the form of a Markov network, into convolutional neural networks has been shown to enhance efficiency compared to purely feedforward models, achieving similar performance in image classification and neural prediction tasks (Zhang et al., 2022). In the context of neural prediction, the underlying assumption is that the closer a model can replicate the neural computation mechanisms responsible for a real neuron's response, the more accurate the model's predictive capabilities become (Pogoncheff et al., 2023; Willeke et al., 2023; Li et al., 2019).

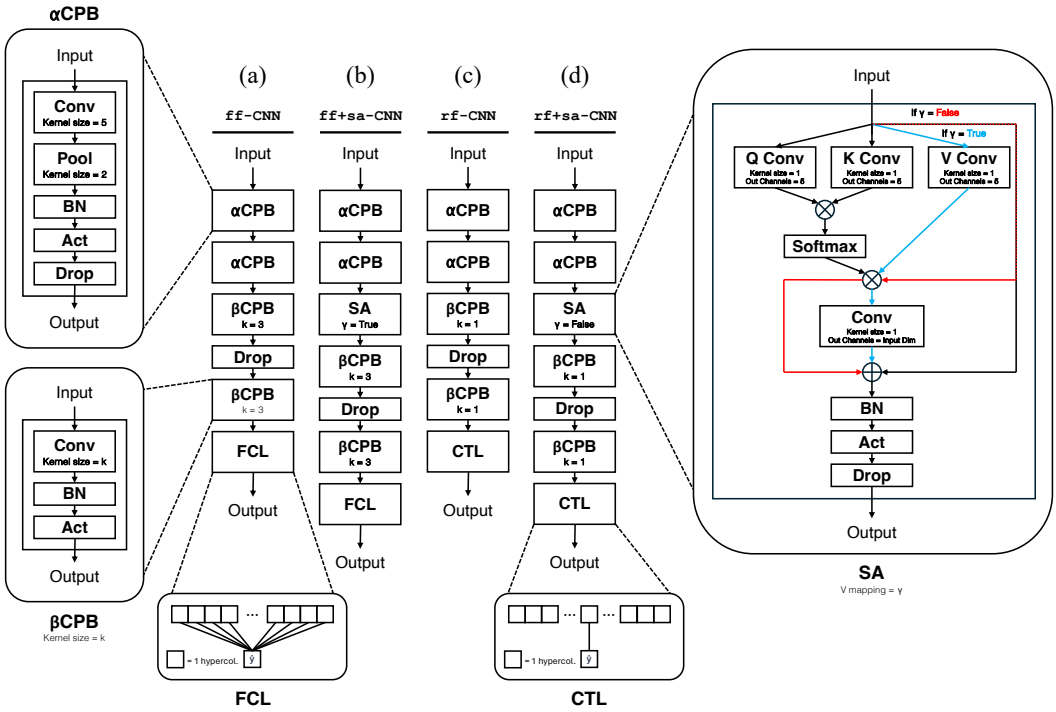


Figure 2: Models explored in this study. Models are constructed from two types of convolutional processing blocks (CPB): αCPB and βCPB. αCPB has a fixed convolution kernel size $= 5$ and max pooling kernel size $= 2$. βCPB takes an input convolution kernel size of k, and has no pooling layers. The two final layer readout modes are fully connected (FCL) and center hypercolumn only (CTL). Self-attention (SA) takes as input a boolean γ that determines whether the value (V) vector is transformed; if $\gamma =$True then V is mapped, otherwise V is equal to the input. All models with SA utilize single-headed attention. **(a)** shows the feedforward CNN. **(b)** shows the feedforward CNN augmented with self-attention. **(c)** shows the receptive field CNN. **(d)** shows the receptive field CNN augmented with self-attention.

Self-attention for global dependencies Self-attention mechanisms have recently become a pivotal component in deep learning models, especially in natural language processing and increasingly in computer vision tasks (Vaswani et al., 2023; Zhao et al., 2020; Kim et al., 2021). In computer vision, self-attention performs a weighted average operation based on the context of input features, computing attention weights dynamically through a similarity function between pixel pairs (Vaswani et al., 2023; Pan et al., 2022). This flexibility allows the attention module to adaptively focus on different regions and capture informative features (Ramachandran et al., 2019). Self-attention has also been integrated with CNNs to enhance their representational power (Pan et al., 2022; Yang et al., 2019; Bello et al., 2020). By enabling CNNs to consider distant spatial relationships within an image, self-attention improves the network's ability to capture global context. This mechanism overcomes the limitations of traditional CNNs, which primarily concentrate on local features because of their convolutional structure. Taking the complementary properties of convolution and self-attention, the benefits of each paradigm can be extracted by integrating the two and using self-attention to augment convolution modules (Dai et al., 2021; Yang et al., 2019; Pan et al., 2022; Cordonnier et al., 2020).

3 Approach

In this study, we developed a set of deep learning models to model V1 neural response to natural images, with the goal of evaluating the potential roles of the self-attention mechanism in neural computation within the visual cortex. We obtained a dataset of neuronal responses measured using two-photon imaging with GCaMP5 from two awake behaving macaque monkeys performing

a fixation task, consisting of 302 neurons from monkey 1 (M1S1) and 299 neurons from monkey 2 (M2S1), in response to 34k and 49k natural images extracted from the ImageNet dataset. The neurons were recorded across five days and tracked anatomically based on landmarks as well as based on their responses to 200 fingerprint images tested every day. The images were presented in sequence with 500 ms per image preceded by 1500 second of blank screen. The 30k-50k images in the training set were presented once, and the 1000 images in the validation set were tested once with 10 repeats. Images were 100×100 pixels, with 30×30 pixels for 1 degree visual angle. The eccentricity of the recording sites were 3 degrees and 1.79 degrees, with average receptive field sizes (diameters at half-height) of about 0.75 and 0.5 degrees, respectively. We preprocessed the dataset before modeling, and notably downsampled input images to 50×50 pixels, yielding 15×15 pixels per degree visual angle. The receptive fields of neurons from each 1 mm $\times$ 1 mm site (the scale/size of a hypercolumn) in macaque monkeys exhibited significant overlap. These fields were mapped using oriented bars or SmoothGrad feature attribution on our deep learning model. The standard deviation of the receptive field centers, all of which are contained within the center hypercolumn of our CNN models, was less than 1 pixel ($1/15$ degree visual angle). See Appendix A.1 for more details of the macaque experimental setup.

3.1 AUGMENTING FEEDFORWARD CNNS WITH SELF-ATTENTION

First, we investigate whether incorporating self-attention into the baseline feedforward CNN model enhances neural response prediction performance. See Appendix A.3 for comparisons to other established models.

Baseline feedforward model (`ff-CNN`) See Figure 2(a) for the architecture. The baseline feedforward model is comprised of two α-convolutional processing blocks (αCPB) and two β-convolutional processing blocks (βCPB), followed by a fully connected readout layer (FCL) . A single `ff-CNN` model is fitted to each neuron. All models described below are derived from this baseline model. Given a grayscale input image with dimensions 50×50 pixels, the two αCPB layers with a 5×5 kernel encode the input of size $(1 \times 50 \times 50)$ into $(c \times 9 \times 9)$ where $c \in \mathbb{N}$ is the number of channels ($c \in \{30, 32\}$ in this study). The center hypercolumn of the post-αCPB encoding has a centered effective receptive field size of 13×13 pixels. In other words, the center hypercolumn of the latent representation after the αCPB layers will have a 13×13 (or 0.8×0.8 degree visual angle) feedforward receptive field at the center of input 50×50 image. Note that the real neurons' receptive fields are contained inside the receptive field of the center hypercolumn. In the baseline model, the two αCPB layers are followed by two βCPB layers with 3×3 kernels to further expand the effective receptive field of the center-hypercolumn. Finally, `ff-CNN` has access to the entirety of the input image in the final layer as the readout has full access to all the hypercolumns. Thus, the baseline `ff-CNN` CNN has two modalities of contextual modulation – convolutions and a fully connected layer.

Feedforward with self-attention model (`ff+sa-CNN`) See Figure 2(b) for the architecture. We augment `ff-CNN` with a self-attention layer immediately after the last αCPB and before the first βCPB. This placement enables SA to act on an adequately convolved feature representation, but also be further modulated by convolutions before feeding into the final layer. We compare the performance of `ff+sa-CNN` against that of `ff-CNN`, controlling the parameter counts to be roughly equal by decreasing the number of channels, which is maintained throughout entire model, from $c = 32$ in the baseline CNNs to $c = 30$ in the self-attention models to account for the addition of the SA layer. In the context of contextual modulation, `ff+sa-CNN` intermixes spatial interactions and inter-channel mixing across SA, the posterior βCPBs, and the FCL.

3.2 FACTORIZING THE CONTEXTUAL MODULATION MECHANISMS

There are three mechanisms in `ff+sa-CNN` mediating contextual interactions. We proceed to factorize `ff+sa-CNN` by removing the contextual modulation contributed by the βCPBs and the FCL to assess the standalone capability of SA in incorporating surrounding context. This is accomplished by constructing a baseline receptive field model and a model where only SA is mediating horizontal connections.

Baseline receptive field model (`rf-CNN`) See Figure 2(c) for the architecture. We first construct the `rf-CNN` model, which is devoid of contextual modulation, by subtracting from `ff-CNN`: the kernel size in the βCPBs are changed from 3×3 to 1×1 and the fully connected layer is changed to look only at the center hypercolumn (CTL). The 1×1 convolutions perform no spatial expansion before feeding into the CTL. Thus, `rf-CNN` is making predictions solely based on the center hypercolumn receptive field produced by the αCPBs, which covers the center 13×13 pixels of the input image.

Receptive field with self-attention model (`rf+sa-CNN`) See Figure 2(d) for the architecture. We add self-attention to `rf-CNN` prior to the βCPBs to construct `rf+sa-CNN`. Self-attention is the **only** mechanism for incorporating surround context in this model. The parameter counts are again controlled by reducing the number of channels from $c = 32$ (in `rf-CNN`) to $c = 30$ (see Appendix A.10 for more details). We compare the performance of the two receptive field models, alongside the feedforward models. Note that in `rf+sa-CNN`, γ is False in the SA layer, meaning SA operates exclusively on the horizontal spatial interactions between hypercolumns without any inter-channel mixing. In contrast, γ is True in the SA layer of `ff+sa-CNN`, which allows channel mixing in SA. Channel mixing potentially provides self-attention greater flexibility (see Appendix A.9 and A.11).

3.3 Incremental Learning: Factorizing the Learning Process

Discussed in Section 4.2, comparing the four models above reveals that contextual modulation introduced in `rf+sa-CNN` via SA did not produce better performance relative to `rf+CNN`, despite `ff+sa-CNN` having clear performance improvements over `ff-CNN`. This is not due to the difference in channel mixing (the γ parameter in SA). We hypothesize that bottle-necking all the gradient signals solely through the center hypercolumn during backpropagation makes it difficult for the network to properly learn the αCPB layers and the SA layer simultaneously. Thus, we investigate an incremental learning paradigm where we allow the receptive fields of the αCPBs to be learned before incorporating any context mechanisms. We then incrementally add and learn a self-attention layer followed by a fully connected readout layer.

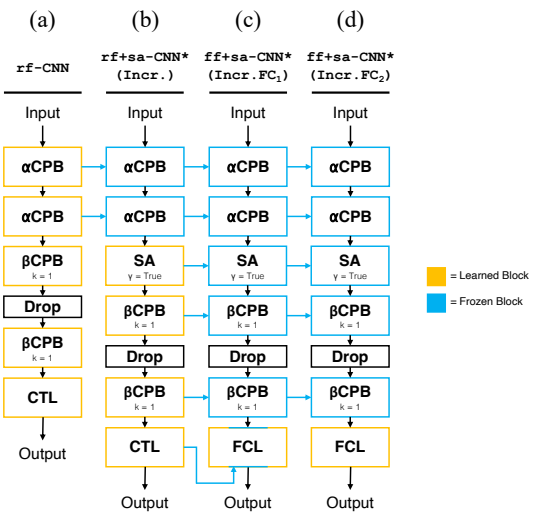


Figure 3: Incremental learning models. **(a)** shows the baseline receptive field CNN, equivalent to Figure 2(c). **(b)** shows (a) augmented with SA and learned incrementally; the αCPBs are taken from (a) and the remaining layers are learned. The * denotes slight modification from `rf+sa-CNN`, Figure 2(d), namely γ is changed to True. **(c), (d)** show the result of replacing the CTL in (b) with a FCL, and learned incrementally; (c) freezes only the center hypercolumn in the FCL ($\texttt{FC}_1$) whereas (d) allows the FCL to learn freely ($\texttt{FC}_2$). (c) and (d) have all other layers taken from (b). The * denotes slight modification from `ff+sa-CNN`, Figure 2(b), namely k in βCPB is changed to $k = 1$. `(Simul.)` models are equivalent in architecture, except all blocks are learned.

The following progression of models, `rf-CNN`, `rf+sa-CNN`*, and `ff+sa-CNN`* (as shown in Figure 3), incrementally expands the capacity of contextual modulation. An important distinction between incremental models and models shown in Figure 2, marked by *, is a 1×1 kernel in the βCPB, which maintains channel mixing but removes further spatial integration through convolution. `rf-CNN` (shown in Figure 3(a) or Figure 2(c)) has information only from the center receptive field. `rf+sa-CNN`* (shown in Figure 3(b)) uses only the self-attention mechanism for contextual modulation. `ff+sa-CNN`* (shown in Figure 3(c)-(d)) has the same surround-center modulation as `rf+sa-CNN`* from self-attention, but allows spatial integration of the global context by changing the CTL to FCL at the end. As hori-

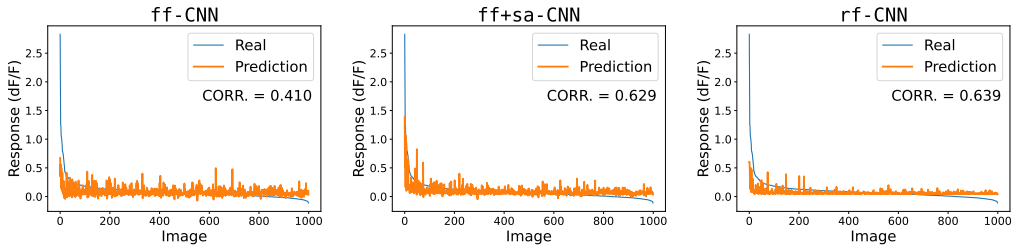


Figure 4: Neuronal tuning curves of `ff-CNN`, `ff+sa-cnn`, and `rf-CNN`. Pearson correlation does not reflect peak tuning. Despite `rf-CNN` having the better correlation, it is clear that `ff+sa-CNN` is able to capture the peak significantly better, at the cost of a noisier overall tuning. Example shown is M1S1 neuron 238. See Appendix A.8 for population averages.

zontal connections in the visual cortex are known to mature after the development of the receptive fields, we designed an incremental learning setup where `rf-CNN` first learns the receptive fields, then `rf+sa-CNN`*`(Incr.)` learns a self-attention layer only after `rf-CNN` has already learned the αCPB receptive fields. Finally, `ff+sa-CNN`*`(Incr.FC`$_1$`)` and `ff+sa-CNN`*`(Incr.FC`$_2$`)` inherit the receptive fields and self-attention structures of `rf+sa-CNN`*, but differ in the change to a FCL readout. Models labelled `(Incr.)` are learned incrementally as such, and models labelled `(Simul.)` are traditionally trained simultaneously.

3.4 Hyperparameter selection and Model training

Rather than splitting the evaluation set for hyperparameter selection, we partitioned our population of neurons to select hyperparameters (training and architectural). We fine-tuned, by experimenting with batch size, learning rate, epochs, number of layers, number of channels per layer, etc., model hyperparameters on a subset of 50 neurons using a relatively coarse grid search. We list key training hyperparameters here: (1) batch size = 50, (2) learning rate = 0.001, (3) optimizer = Adam, (4) loss = MSE, (5) epochs = 50. Training and computations were performed on an in-house computing cluster with GPU (NVIDIA V100 or similar) nodes. Training hyperparameters were held constant across all models. Architectural hyperparameters were held constant across layers shared between models. We do not optimize hyperparameters for models other than the baseline `ff-CNN`.

The primary objective of this project is to demonstrate that self-attention can enhance neural response prediction relative to the baseline feedforward CNN, despite hyperparameters being optimized only for the baseline model. Since we show that `ff+sa-CNN` improves upon `ff-CNN` in both evaluation metrics (see Section 4.1), further hyperparameter optimization is unnecessary for our objective. Instead, we are interested in understanding the reason behind this improvement. The other models tested in this study are architectural subsets of the `ff+sa-CNN`, designed to dissect their contributions to its success. We do not anticipate any derivative models outperforming `ff+sa-CNN`, justifying holding hyperparameters constant across models for fair comparison.

3.5 Model evaluation metrics

To quantify performance, models were evaluated on two criteria, Pearson correlation and peak tuning index. Pearson correlation represents the overall tuning similarity between a model's predicted responses and the real neuron's recorded responses. The peak tuning index is used to quantify how well a model can predict and match in magnitude the strongest responses recorded by the real neuron. This lets us evaluate how well a model can discriminate between, as well as model the response magnitude of, images that are strongly excitatory and images that incite a weak response.

Pearson correlation: The Pearson correlation (CORR.) is taken between neuron responses and model-predicted responses. Pearson correlation is a standard measure for evaluating neural response prediction. Other established measures, including FEV, r^2_{er}, and CC^2_{norm} (Willeke et al., 2022; Cadena et al., 2017; Pospisil & Bair, 2021; Zhang et al., 2022), were used as well and yielded similar results to Pearson correlation (see Appendix A.3 and A.4).

In this paper, we explored a set of measures to assess the peak of a neuronal tuning curve. In our macaque V1 dataset, we found that neurons exhibit sharp stimulus selectivity, consistent with findings from Tang et al. (2018a;b), reinforcing the diversity and complexity of V1 neurons (see Appendix A.2). We found that Pearson correlation and other standard metrics (see Appendix A.4) are successful in measuring a model's fit to the overall tuning curve, but often fail to represent the peak tuning preference of neurons, of which is a key aspect of a neuronal function. For example, Figure 4 demonstrates that while `rf+CNN` achieves a higher correlation in approximating the overall tuning curve compared to `ff+sa-CNN`, the latter outperforms in modeling the neuron's tuning curve peak. To address this issue, we developed two new metrics to better assess a model's ability to capture the peak tuning of neurons.

Peak tuning index: The peak tuning index (PT) is a membership metric of the strongest predictions above a threshold determined by the top 1% of real responses. PT can be roughly interpreted as the percentage of the peak that a model captures, under a magnitude prior. The index is calculated as:

$$\text{PT} = \frac{\texttt{\# of top 1\% predictions} \geq \texttt{min(top 1\% real responses)}}{\texttt{\# of responses in the top 1\%}} \times 100\%$$

PT is divided into PT_J and PT_S, based on how `# of top 1% predictions` is defined. PT_J is when predictions are jointly rank ordered with respect to the real responses. PT_S is when predictions are separately rank ordered independently of the real responses. PT_J is a stricter measure. Note that because we train with MSE loss, models are incentivized to minimize the absolute difference between predictions and real responses, rather than match the curvature of the tuning curve. This minimizes the risk of PT being misrepresentative due to lateral shifts in the tuning curve.

4 RESULTS

4.1 SELF-ATTENTION IMPROVES NEURAL RESPONSE PREDICTION

Table 1: Average Pearson correlation and peak tuning metrics for models trained on M1S1 and M2S1. Correlation SEM $= 0.009$ was consistent across models and monkeys. Despite `rf-CNN` unexpectedly outperforming `rf+sa-CNN`, the difference is recovered when `rf+sa-CNN` is trained incrementally (see Section 4.2).

	M1S1				M2S1			
Model	CORR.	Δ `ff-CNN`	PT_J	PT_S	CORR.	Δ `ff-CNN`	PT_J	PT_S
`ff-CNN`	0.393	0.0%	3.3 ± 0.5	5.6 ± 0.9	0.477	0.0%	8.6 ± 0.9	16.2 ± 1.6
`ff+sa-CNN`	**0.416**	+6.6%	$\mathbf{5.6} \pm 0.6$	$\mathbf{10.5} \pm 1.1$	**0.491**	+3.3%	$\mathbf{11.5} \pm 0.9$	$\mathbf{23.5} \pm 1.8$
`rf-CNN`	**0.420**	+8.6%	$\mathbf{1.1} \pm 0.3$	$\mathbf{1.8} \pm 0.5$	**0.496**	+4.3%	$\mathbf{4.4} \pm 0.6$	$\mathbf{6.6} \pm 1.0$
`rf+sa-CNN`	0.414	+7.2%	0.7 ± 0.2	1.0 ± 0.3	0.486	+2.4%	3.4 ± 0.5	5.1 ± 0.8

We compared the performance of the `ff+sa-CNN` model to the parameter-matched baseline `ff-CNN` model and found that incorporating self-attention significantly improved correlation and both peak tuning metrics (see first two rows of Table 1). This indicates that self-attention enhances modeling of both the overall tuning and peak tuning aspects of the neurons, with consistent results across both monkeys.

It is important to note that the `ff-CNN` center hypercolumn at the readout layer has a receptive field much larger than the real neuron's receptive field, due to successive convolutions in the αCPBs and βCPBs. Additionally, the fully connected readout layer also incorporates long-range spatial dependencies. Thus, the self-attention layer in `ff+sa-CNN` acts as an additional mechanism for modeling horizontal connections, and provides additional performance benefits.

To better understand the role of self-attention in contextual modulation, we constructed a baseline receptive field model, `rf-CNN`, that is devoid of all contextual modulation mechanisms. `rf-CNN`'s CTL readout only uses the center-hypercolumn of the convolved feature space to make predictions. Note that the feedforward receptive field of the center hypercolumn after the two αCPBs is a centered 13×13 pixel portion of the input image. Moreover, between the αCPB layers and the CTL layer,

the size of the center hypercolumn receptive field does not expand due the the 1×1 kernel in the βCPBs. This means that predictions are being made solely based on the center 13×13 pixels, which corresponds roughly to the real neuron's receptive field.

Surprisingly, we found that `rf-CNN` achieved the highest correlation, indicating that models focusing primarily on the classical receptive field offer the best fit to the overall tuning curve (see Table 1). However, correlation fails to reflect the model's shortcomings in fitting the peak of the tuning curve. Table 1 also reveals that while `rf-CNN` has the highest Pearson correlation, it performs worse at capturing the tuning peak compared to `ff+sa-CNN` and `ff-CNN`. Thus, we conclude that contextual modulation plays a crucial role in peak tuning, and that the three mechanisms in `ff+sa-CNN` for integrating surround information are complementary.

4.2 Dissecting Contextual Modulation Mechanisms via Incremental Learning

We explore the relative contributions of the different contextual modulation mechanisms. Specifically, we first ask: Is self-attention alone sufficient to model contextual modulation? To answer this question, we added SA to `rf-CNN` to produce `rf+sa-CNN`. We found that without posterior βCPBs and a FCL, the self-attention in `rf+sa-CNN` is not useful. In fact, the performance is worse than `rf-CNN` in overall correlation and both peak tuning metrics (see Table 1). This result is somewhat unexpected, as the addition of self-attention in `ff+SA-CNN` does improve upon `ff-CNN`. It is possible that the spatial integration mechanisms in the βCPBs, along with the FCL, are necessary to provide sufficient pathways for backpropagating the gradients during learning, so that both the αCPB receptive fields and self-attention kernels can be properly learned.

To test this hypothesis, we explored whether an incremental learning approach, where different network components are learned sequentially, could yield a model that performs on par with `ff+sa-CNN`. We show that learning the feedforward kernels in the αCPBs first, followed by learning the self-attention layer, and finally the fully connected layer, can nearly match the performance of `ff+sa-CNN`. This indicates that although the spatial integration by βCPBs contributes to performance, the self-attention layer plays a more critical role in capturing horizontal interactions crucial for modeling peak tuning.

Table 2: Average Pearson correlation for models incrementally and simultaneously trained on M1S1 and M2S1. Correlation SEM $= 0.009$ was consistent across models and monkeys.

	M1S1		M2S1	
Model (Training Method)	CORR.	Δ `rf-CNN`	CORR.	Δ `rf-CNN`
`rf-CNN(Simul.)`	**0.420**	0.0%	**0.496**	0.0%
`rf+sa-CNN*(Simul.)`	0.409	−2.6%	0.480	−3.2%
`rf+sa-CNN*(Incr.)`	**0.421**	+0.6%	**0.493**	−0.3%
`ff+sa-CNN*(Simul.)`	0.416	−0.8%	0.490	−0.7%
`ff+sa-CNN*(Incr.FC1)`	**0.430**	+3.0%	**0.494**	−0.1%
`ff+sa-CNN*(Incr.FC2)`	0.414	−1.3%	0.488	−1.1%

Incremental learning offers a valuable approach for accurately assessing the potential of each spatial integration mechanism in modeling contextual modulation and evaluating their interdependence in generating an effective model. Table 2 highlights several incremental models we tested and their relative improvements. Learning the receptive field first, followed by learning self-attention in `rf+SA-CNN*(Incr.)`, outperforms learning both simultaneously in `rf+SA-CNN*(Simul.)` (see second and third rows of Table 2, and middle pairs in Figure 5). This supports our hypothesis that jointly learning the αCPB receptive fields and self-attention may overwhelm the system when gradients are constrained to propagate through the center hypercolumn alone. Incremental learning can improve `rf+sa-CNN*` to match the peak prediction performance of `rf-CNN`, but not beyond. This suggests that, when used with a CTL readout, self-attention alone is insufficient to fully capture peak tuning.

Table 2 and Figure 5 demonstrate that removing the CTL restriction–i.e., allowing the readout to access information from the hypercolumns in the final convolutional layer via FCL, as in `ff+sa-CNN*`–enables the network to nearly match the performance of the `ff+sa-CNN`. `ff+sa-CNN*` is named as such because it closely resembles `ff+sa-CNN`, with the only differ-

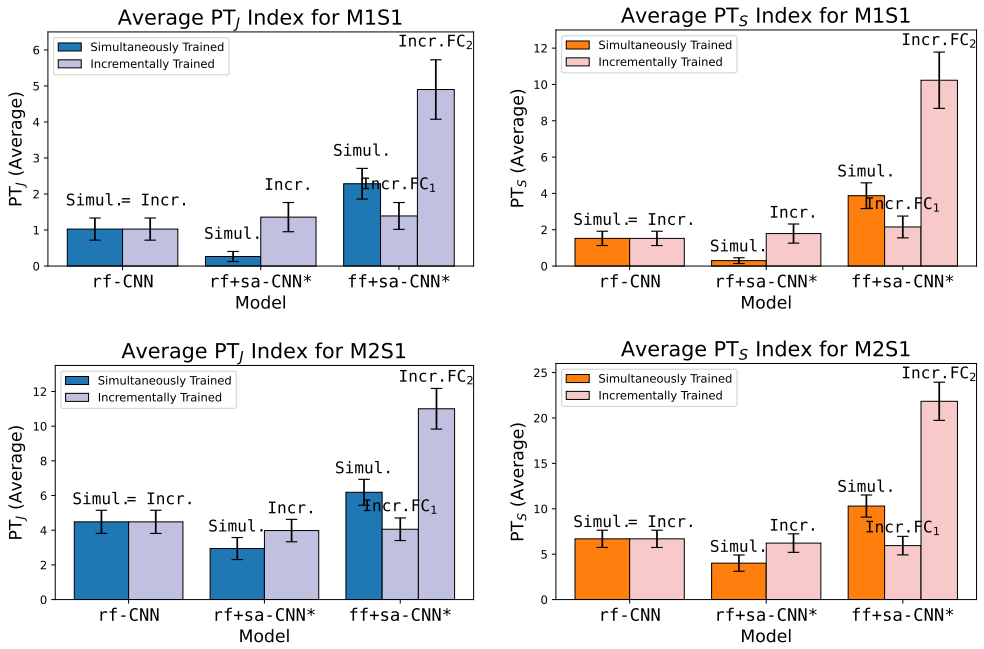


Figure 5: Average peak tuning indices for incrementally and simultaneously trained models. **Top row:** bar charts for M1S1. **Bot row:** bar charts for M2S1. **Left col:** average PT$_J$ values. **Right col:** average PT$_s$ values. Error bars are SEM.

ence being the use of a 1×1 kernel instead of a 3×3 kernel in the βCPBs. These findings suggest that posterior convolution is not required for spatial integration after the self-attention layer when the model is trained incrementally.

A reoccurring observation is that focusing on receptive field information tends to improve overall correlation, while emphasizing contextual information enables the network to better model peak tuning. This pattern is evident when comparing `rf-CNN` with `ff+sa+CNN` in Table 1. A similar contrast exists between `ff+sa-CNN*(Incr.FC`$_1$`)` and `ff+sa-CNN*(Incr.FC`$_2$`)`. In `ff+sa-CNN*(Incr.FC`$_1$`)`, the model inherits the center hypercolumn weights and then learns the surrounding hypercolumn contributions through the FCL. In contrast, `ff+sa-CNN*(Incr.FC`$_2$`)` learns the weights of all hypercolumns in the FCL simultaneously. While the former excels in correlation, the latter performs better in peak tuning. This suggests that the receptive field is most important towards overall tuning, whereas surround-center interactions are key to capturing peak tuning.

The `ff+sa-CNN*` models saw an improvement in PT$_J$ and PT$_S$ over the `rf+sa-CNN*` models. This suggests that either the fully connected layer (FCL) plays a critical role in predicting peak responses, or that constraining the readout to the center hypercolumn (CTL) in `rf+sa-CNN*` limits error propagation to the self-attention block during training. As a result, self-attention may be inadequately learned in these cases, impairing the model's ability to effectively implement contextual modulation.

4.3 Incremental learning emphasizes the contribution of the classical receptive field

A well-established neurophysiological principle is that stimuli within the classical receptive field of V1 neurons are the primary driver of neural responses, while the contextual surround modulates them. We found that when the entire network is trained simultaneously, as in `ff+sa-CNN*(Simul.)`, performance is weaker and the network does not follow this principle. However, with incremental learning, the center hypercolumn develops into the dominant contributor, as observed in `ff+sa-CNN*(Incr.FC`$_1$`)` and `ff+sa-CNN*(Incr.FC`$_2$`)`. Figure 6 illus-

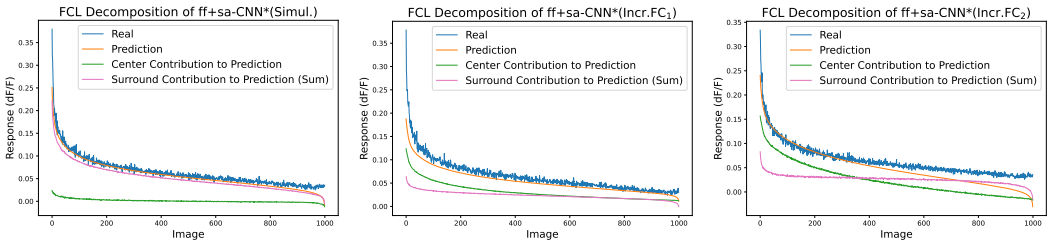


Figure 6: Average FCL decomposition of `ff+sa-CNN*` when trained differently. The center contribution (green) and the total surround contribution (pink) sum to the prediction tuning curve (orange). Plots are rank ordered with respect to predicted responses. Averages are calculated by plotting rank ordered decomposed tuning curve for each neuron, then averaging across each image. Individual contributions from hypercolumns can be found in Appendix A.5.

trates the sum of connection weights from the readout, showing that in `ff+sa-CNN*(Simul.)`, weights are evenly distributed, whereas in the incrementally trained models, the center hypercolumn, corresponding to the classical receptive field, is emphasized. Incremental learning fosters this center-surround division and modulation, yielding interpretable performance benefits.

5 Discussion

CNNs are widely used and effective models for visual cortical neurons, and they inherently include two mechanisms for contextual modulation: successive convolutions and a fully connected layer, which allow the global image context to be accessible to the readout. In this paper, we demonstrated that augmenting CNN models of cortical neurons with self-attention enhances predictions of both the overall tuning curve and the tuning peak. Self-attention, resembling three-way interactions in probabilistic graphical models, facilitates flexible center-surround interaction via contextual variables (Coen-Cagli et al., 2015; Fei & Pitkow, 2022). This provides additional flexibility and complementary benefits to the CNN's inherent context mechanisms. While large-scale transformer models with multiple attention heads have achieved state-of-the-art performance in modeling mouse V1 neurons by capturing long-range dependencies (Li et al., 2023), our work isolates the contributions of self-attention in CNNs for modeling horizontal circuits, highlighting the dependencies and complementarity among three different mechanisms of contextual modulation.

Several key findings emerged from this work that advance our understanding of cortical computation and neural codes. First, we found that focusing on receptive field information, as in `rf-CNN`, yields the highest Pearson correlation, alongisde other standard measures (see Appendix A.4), for overall neuronal tuning curves (see Table 1). This suggests that the classical receptive field is the primary driver behind a neuron's overall response. Our incremental learning experiments further reinforce the importance of focusing first on the bottom-up information within the classical receptive field before learning the horizontal connections (see Figure 6). Second, we demonstrated that contextual modulation is important for predicting the peak response, with self-attention playing a pivotal role. A trade-off, however, exists between the receptive field and surround modulation: RF-centric models fit overall tuning curves more accurately, while increased contextual surround modulation enhances peak tuning, though often at the expense of overall tuning correlation. Incremental learning, which fosters a strong receptive field bias, may help even out this trade-off. This is consistent with neurophysiological evidence supporting a dominant classical receptive field and weaker surround modulation, with recurrent connections being fine-tuned after receptive field development.

A recent CNN-based model of mouse V1 neurons revealed that the most excitable images often involve stimulus features outside the receptive fields that are consistent with the concept of pattern completion (Fu et al., 2023). Similarly, we found that models capable of capturing peak tuning display interpretable contextual modulation, such as association fields and pattern completion, within the self-attention module (see Appendix A.6). Additionally, incorporating a self-attention layer improved models' data efficiency (see Appendix A.7). Further theoretical and experimental investigations are needed to characterize and evaluate the interactions facilitated by self-attention, so as to uncover how these mechanisms may be implemented by biological circuits.

6 ETHICS STATEMENT

All procedures involving animals for generating the data of this paper were in accordance with the Guide for the Institutional Animal Care and Use Committee (IACUC) of Peking University.

7 ACKNOWLEDGMENTS

This work was supported by NSF CISE RI 1816568 and NIH R01 EY030226-01A1 awarded to Tai Sing Lee. Isaac Lin is supported by the NSF REU supplement of RI 1816568. Imaging data was produced by Tianye Wang and Shiming Tang with the support of STI2030-Major Projects 2022ZD0204600, National Natural Science Foundation of China U1909205, and funds from the Peking-Tsinghua Center for Life Sciences to ST.

REFERENCES

Alessandra Angelucci and Paul C. Bressloff. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate v1 neurons. In S. Martinez-Conde, S.L. Macknik, L.M. Martinez, J.-M. Alonso, and P.U. Tse (eds.), *Visual Perception*, volume 154 of *Progress in Brain Research*, pp. 93–120. Elsevier, 2006. doi: https://doi.org/10.1016/S0079-6123(06)54005-1. URL https://www.sciencedirect.com/science/article/pii/S0079612306540051.

Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks, 2020.

Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. 2017. doi: 10.1101/201764.

James Cavanaugh, Wyeth Bair, and J Movshon. Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *Journal of neurophysiology*, 88:2530–46, 12 2002. doi: 10.1152/jn.00692.2001.

Ruben Coen-Cagli, Adam Kohn, and Odelia Schwartz. Flexible gating of contextual influences in natural vision. *Nature neuroscience*, 18, 10 2015. doi: 10.1038/nn.4128.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers, 2020.

R. James Cotton, Fabian H. Sinz, and Andreas S. Tolias. Factorized neural processes for neural processes: k-shot prediction of neural responses, 2020.

Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes, 2021.

Yicheng Fei and Xaq Pitkow. Attention as inference with third-order interactions. In *NeurIPS '22 Workshop on All Things Attention: Bridging Different Perspectives on Attention*, 2022. URL https://openreview.net/forum?id=s2VfopqfA0.

Daniel J. Felleman and David C. van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1 1:1–47, 1991. URL https://api.semanticscholar.org/CorpusID:9334496.

Jiakun Fu, Suhas Shrinivasan, Kayla Ponder, Taliah Muhammad, Zhuokun Ding, Eric Wang, Zhiwei Ding, Dat T. Tran, Paul G. Fahey, Stelios Papadopoulos, Saumil Patel, Jacob Reimer, Alexander S. Ecker, Xaq Pitkow, Ralf M. Haefner, Fabian H. Sinz, Katrin Franke, and Andreas S. Tolias. Pattern completion and disruption characterize contextual modulation in mouse visual cortex. *bioRxiv*, 2023. doi: 10.1101/2023.03.13.532473. URL https://www.biorxiv.org/content/early/2023/03/14/2023.03.13.532473.

Kuan Han, Haiguang Wen, Yizhen, Di Fu, Eugenio Culurciello, and Zhongming Liu. Deep predictive coding network with local recurrent processing for object recognition, 2018.

Kyungmin Kim, Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Zhicheng Yan, Peter Vajda, and Seon Kim. Rethinking the self-attention in vision transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3065–3069, 2021. doi: 10.1109/CVPRW53098.2021.00342.

William F. Kindel, Elijah D. Christensen, and Joel Zylberberg. Using deep learning to probe the neural code for images in primary visual cortex. *Journal of Vision*, 19(4):29–29, 04 2019. ISSN 1534-7362. doi: 10.1167/19.4.29. URL https://doi.org/10.1167/19.4.29.

David A. Klindt, Alexander S. Ecker, Thomas Euler, and Matthias Bethge. Neural system identification for large populations separating "what" and "where", 2018.

Nikolaus Kriegeskorte. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1 (Volume 1, 2015):417–446, 2015. ISSN 2374-4650. doi: https://doi.org/10.1146/annurev-vision-082114-035447. URL https://www.annualreviews.org/content/journals/10.1146/annurev-vision-082114-035447.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. Brain-like object recognition with high-performing shallow recurrent anns, 2019.

Bryan M. Li, Isabel M. Cornacchia, Nathalie L. Rochefort, and Arno Onken. V1t: large-scale mouse v1 response prediction using a vision transformer, 2023. URL https://arxiv.org/abs/2302.03023.

Ming Li, Fang Liu, Hongfei Jiang, Tai Sing Lee, and Shiming Tang. Long-term two-photon imaging in awake macaque monkey. *Neuron*, 93(5):1049–1057.e3, 2017. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2017.01.027. URL https://www.sciencedirect.com/science/article/pii/S089662731730051X.

Zhe Li, Wieland Brendel, Edgar Y. Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian H. Sinz, Xaq Pitkow, and Andreas S. Tolias. Learning from brains how to regularize machines, 2019.

Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay Jagadish, Eric Wang, Edgar Y. Walker, Santiago A Cadena, Taliah Muhammad, Erick Cobos, Andreas S. Tolias, Alexander S Ecker, and Fabian H. Sinz. Generalization in data-driven models of primary visual cortex. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Tp7kI90Htd.

Nikola T. Markov, Julien Vezoli, Pascal Chameau, Arnaud Falchier, René Quilodran, Cyril Huissoud, Camille Lamy, Pierre Misery, Pascale Giroud, Shimon Ullman, Pascal Barone, Colette Dehay, Kenneth Knoblauch, and Henry Kennedy. Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology*, 522(1):225–259, 2014. doi: https://doi.org/10.1002/cne.23458. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.23458.

Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J. DiCarlo, and Daniel L. K. Yamins. Task-driven convolutional recurrent models of the visual system, 2018.

Xu Pan, Annie DeForge, and Odelia Schwartz. Generalizing biological surround suppression based on center surround similarity via deep neural network models. *bioRxiv*, 2023. doi: 10.1101/2023.03.18.533295. URL https://www.biorxiv.org/content/early/2023/03/19/2023.03.18.533295.

Xuran Pan, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. On the integration of self-attention and convolution, 2022.

Galen Pogoncheff, Jacob Granley, and Michael Beyeler. Explaining v1 properties with a biologically constrained deep learning architecture, 2023.

Dean A. Pospisil and Wyeth Bair. The unbiased estimation of the fraction of variance explained by a model. *PLOS Computational Biology*, 17(8):1–36, 08 2021. doi: 10.1371/journal.pcbi.1009212. URL https://doi.org/10.1371/journal.pcbi.1009212.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models, 2019.

Kosei Sasaki, Elizabeth C. Cropper, Klaudiusz R. Weiss, and Jian Jing. Functional differentiation of a population of electrically coupled heterogeneous elements in a microcircuit. *Journal of Neuroscience*, 33(1):93–105, 2013. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.3841-12.2013. URL https://www.jneurosci.org/content/33/1/93.

S. Shushruth, Lauri Nurminen, Maryam Bijanzadeh, Jennifer M. Ichida, Simo Vanni, and Alessandra Angelucci. Different orientation tuning of near- and far-surround suppression in macaque primary visual cortex mirrors their tuning in human perception. *The Journal of Neuroscience*, 33: 106 – 119, 2013. URL https://api.semanticscholar.org/CorpusID:13650720.

Courtney J Spoerer, Tim C Kietzmann, Johannes Mehrer, Ian Charest, and Nikolaus Kriegeskorte. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *bioRxiv*, 2020. doi: 10.1101/677237. URL https://www.biorxiv.org/content/early/2020/08/21/677237.

Shiming Tang, Yimeng , Zhihao Li, Ming Li, Fang Liu, Hongfei Jiang, and Tai Sing Lee. Large-scale two-photon imaging revealed super-sparse population codes in the v1 superficial layer of awake monkeys. *eLife*, 7:e33370, apr 2018a. ISSN 2050-084X. doi: 10.7554/eLife.33370. URL https://doi.org/10.7554/eLife.33370.

Shiming Tang, Tai Sing Lee, Ming Li, Yimeng, Yue Xu, Fang Liu, Benjamin Teo, and Hongfei Jiang. Complex pattern selectivity in macaque primary visual cortex revealed by large-scale two-photon imaging. *Current Biology*, 28(1):38–48.e3, 2018b. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2017.11.039. URL https://www.sciencedirect.com/science/article/pii/S096098221731521X.

Gido van de Ven, Tinne Tuytelaars, and Andreas Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4:1–13, 12 2022. doi: 10.1038/s42256-022-00568-3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Hansel, Christoph Blessing, Konstantin-Klemens Lurz, Max F. Burg, Santiago A. Cadena, Zhiwei Ding, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Kaiwen Deng, Yuanfang Guan, Yiqin Zhu, Kaiwen Xiao, Xiao Han, Simone Azeglio, Ulisse Ferrari, Peter Neri, Olivier Marre, Adrian Hoffmann, Kirill Fedyanin, Kirill Vishniakov, Maxim Panov, Subash Prakash, Kishan Naik, Kantharaju Narayanappa, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. Retrospective on the sensorium 2022 competition. In Marco Ciccone, Gustavo Stolovitzky, and Jacob Albrecht (eds.), *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pp. 314–333. PMLR, 28 Nov–09 Dec 2022. URL https://proceedings.mlr.press/v220/willeke23a.html.

Konstantin F. Willeke, Kelli Restivo, Katrin Franke, Arne F. Nix, Santiago A. Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S. Ecker, Fabian H. Sinz, and Andreas S. Tolias. Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. *bioRxiv*, 2023. doi: 10.1101/2023.05.12.540591. URL https://www.biorxiv.org/content/early/2023/05/13/2023.05.12.540591.

Daniel Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19:356–365, 2016. URL https://api.semanticscholar.org/CorpusID:16970545.

Baosong Yang, Longyue Wang, Derek Wong, Lidia S. Chao, and Zhaopeng Tu. Convolutional self-attention networks, 2019.

Yimeng Zhang, Tai Sing Lee, Ming Li, Fang Liu, and Shiming Tang. Convolutional neural network models of v1 responses to complex patterns. *Journal of Computational Neuroscience*, 46:33 – 54, 2018. URL https://api.semanticscholar.org/CorpusID:46934991.

Yimeng Zhang, Harold Rockwell, Sicheng Dai, Ge Huang, Stephen Tsou, Yuanyuan Wei, and Tai Sing Lee. Recurrent networks improve neural response prediction and provide insights into underlying cortical circuits, 2022. URL https://arxiv.org/abs/2110.00825.

Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition, 2020.

A APPENDIX

A.1 ADDITIONAL DETAILS ON MACAQUE EXPERIMENTAL SETUP

We collected data using a nearly identical experimental protocol as detailed in (Tang et al., 2018a;b), except that our dataset was considerably larger, including up to 34K and 49K stimulus-response pairs for the two monkeys respectively.

During each fixation task, a blank screen was presented for 1500 ms after the monkey established fixation, followed by the presentation of a visual stimulus for 500 ms. 34,000 stimuli were tested in monkey 1 and 49,000 stimuli were tested in monkey 2 over a data collection period of 5 days.Neurons were registered anatomically and also by testing a 200 stimuli finger-print each day. To quantify the neural response, a differential image of GCamp5s calcium signals between the stimulus period and blank period was computed for each trial. The dF/F was then calculated based on a 200 ms to 600 ms window after stimulus onset. We note that GCamp5s is slow, but the signal has been found to be correlated with firing rate (Li et al., 2017). The training set was collected with 1 repeat for each stimulus. The validation set consists of 1000 stimuli, each with 10 repeats.

A.2 V1 PREFERRED FEATURES

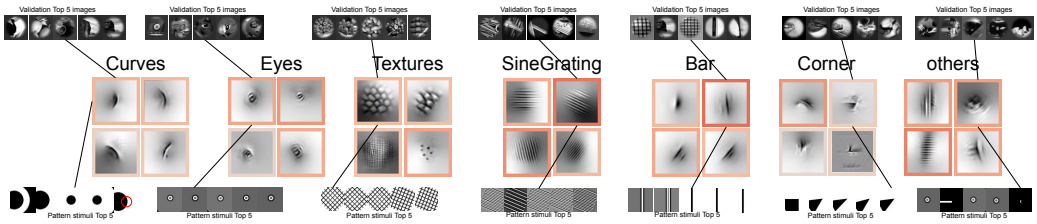
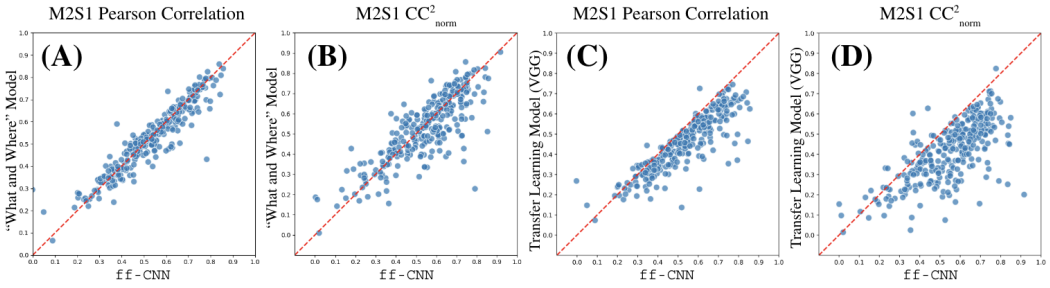


Figure 7: V1 neuron exhibit diversity and complexity in preferred features. Individual CNNs were used to model neurons' responses to a large set of natural images. **Mid row:** visualizations of the optimal stimuli for `ff-CNN` models for 28 neurons. Neurons are clustered into 7 equally sized classes representing preferences for curves, rings (eyes), textures, grating, bars, corners, and other more complex higher order features. **Top row:** top 5 natural images in the validation set that elicited the largest response from real neurons in each class. **Bot row:** top 5 artificial stimuli that elicited the largest response from `ff-CNN` models of neurons in each class. The artificial stimuli preferences are consistent with the optimal stimuli seen in the middle row.

A driving motivation of our study was to identify the natural image features that V1 neurons prefer, which include corners, curvatures, junctions, rings and other higher order features, rather than the traditional orientation and frequency tunings. While traditional artificial stimuli produce a tuning curve, these stimuli rarely represent the neuron's most preferred stimulus. The studies referenced above show that the neurons' true preferences are not necessarily a specific orientation or spatial frequency, and may not be well modeled by Gabor functions and traditional linear/nonlinear models. Rather, the preferred features, which are encoded by the peak responses of neurons as visualized in Figure 7, exhibited diversity and complexity in the tuned features. The objective of this study is not to prove or argue for the diversity and complexity of V1 neural codes, but to explore the mechanisms contributing to them. A key finding of this paper is that the surround contextual mechanism, implemented by a non-local network (i.e. self-attention), positively contributes to the generation of the response peak, and is related to the encoding of preferred higher-order features in neurons. As such,this study aims to model the peak response of neurons and develop a novel metric to evaluate the model's accuracy in capturing the peak of each neuron's "natural image tuning curve."

A.3 COMPARISONS TO OTHER ESTABLISHED MODELS



Hyperparameter Changes		CORR.
ff-CNN with:	N/A (original ff-CNN)	**0.444**
	L1 regularization in readout layer	0.401
	Poisson loss during training	0.409
	All dropout layers removed	0.410

(E)

Model	PT_J	PT_S
ff-CNN	**0.159** ± 0.023	**0.290** ± 0.041
"What and where" Model	0.100 ± 0.024	0.210 ± 0.045

(F)

Figure 8: Comparing the performance of ff−CNN. **(A), (C)** shows the Pearson correlation comparisons of ff−CNN with the "what and where" (shared core factorized) model and the transfer learning (goal driven VGG) model, respectively. **(B), (D)** show CC norm squared comparisons between models. **(E)** shows hyperparameter experiments on a subset of 50 neurons from M2S1. **(F)** Comparison between ff−CNN and the "what and where" model on the ability to capture M2S1 tuning peaks. All results on M1S1 are similar.

Four major classes of models are found in neural response prediction literature: (1) transfer learning models, (2) single CNN feedforward models, the (3) shared core factorized ("what and where") model (Klindt et al., 2018; Lurz et al., 2021), and more recently a (4) transformer based model for mouse V1. The first three types of models have been used in macaque V1, whereas the transformer based model has only been used in mouse V1. We experimented with the first three classes of models and found that for our dataset, the performance of our baseline feedforward models (ff−CNN) are comparable, and at times better, to the transfer learning model and the shared core factorized mode (see Figure 8). Thus, we used the single CNN feedforward model as our baseline model. We did not compare our model with transformer based models such as ViT (Li et al., 2023), which use deep and complex layers to achieve SOTA neural response prediction. The focus of our research and our contributions are different. We demonstrate that surround contextual modulation is critical in predicting the peak responses of macaque V1 neurons in response to natural images. Towards this, we used self-attention to model horizontal interactions, rather than just using an entire transformer module to achieve SOTA performance.

A.4 COMPARISONS TO OTHER ESTABLISHED METRICS

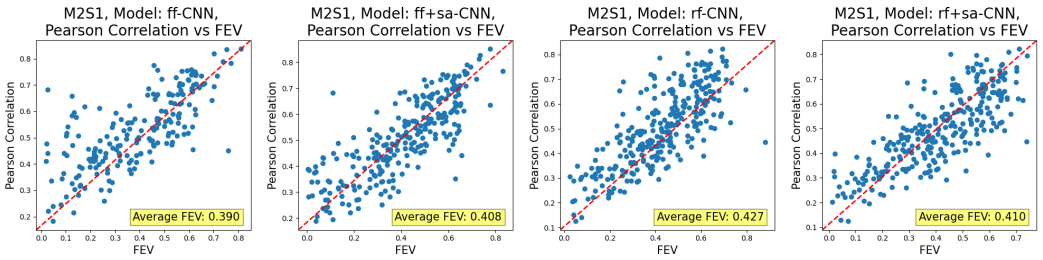


Figure 9: Pearson correlation versus FEV scatter plot for different models. Average FEV is calculated by averaging over neurons in M2S1. Note that the performance of models improve relative to the `ff-CNN` baseline when evaluated based on FEV, following a similar trend to that of Pearson correlation.

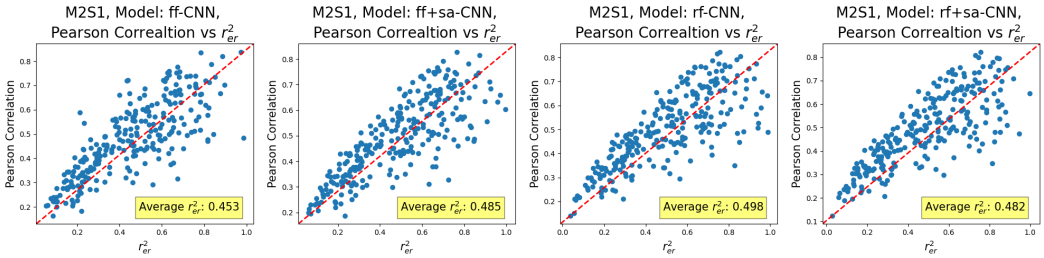


Figure 10: Pearson correlation versus r^2_{er} scatter plot for different models. Average r^2_{er} is calculated by averaging over neurons in M2S1. Note that the performance of models improve relative to the `ff-CNN` baseline when evaluated based on r^2_{er}, following a similar trend to that of Pearson correlation.

We compute measures FEV and r^2_{er} for each of our models. We calculated the average FEV and r^2_{er} across all neurons in M2S1 for different models. Similar to Pearson correlation and CC^2_{norm}, FEV and r^2_{er} measures are similar across models used in our study, and are comparable to established baselines (see Figure 9 and Figure 10). Despite FEV and r^2_{er} taking prediction magnitudes into account, they are still unable to capture the peak tuning properties. This can be attributed to the high degree of sparsity in individual neurons' tuning curve (to natural stimuli), where only 0.4% of the stimuli above half height on average. Because standard measures of performance are heavily influenced by remaining 99% low-responding stimuli, they are not sufficient for capturing the sharp tuning curves peaks we observed (i.e.recognizing the most preferred stimuli of the neurons).

A.5 FCL DECOMPOSITION: AVERAGE CONTRIBUTION FROM EACH HYPERCOLUMN

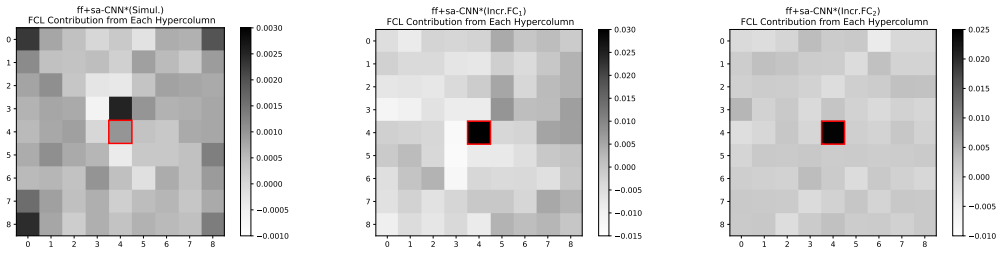


Figure 11: Average FCL contribution from each hypercolumn for M1S1. Center hypercolumn outlined in red. Each heatmap is independently contrast normalized. Note that the distribution in the first heatmap is evenly distributed compared to the others (see contrast scale).

Figure 6 plots the individual contribution of the center hypercolumn with the sum of all surrounding hypercolumns. We observed an evenly distributed contribution from all hypercolumns in the simultaneously trained model, but a strong center contribution in incrementally learned models. This effect is further observed when we display the average contribution of each hypercolumn (see Figure 11).

A.6 MODELS THAT CAN CAPTURE PEAK TUNING EXHIBIT INTERPRETABLE CONTEXTUAL MODULATION SUCH AS ASSOCIATION FIELDS IN THE SELF-ATTENTION MODULE

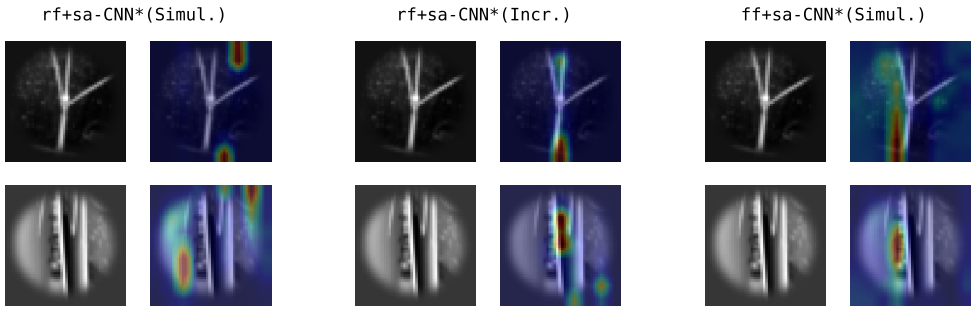


Figure 12: Attention highlighting for incremental learning models. Top two highest response inducing image are shown for M1S1 neuron 153. Note that `rf+sa-CNN*(Incr)`, `ff+sa-CNN*(Incr.FC1)`, and `ff+sa-CNN*(Incr.FC2)` all have the same attention map due to the freezing scheme. The center hypercolumn is queried for highlighted images.

Comparing the attention highlighting between self-attention models from the incremental learning experiment, we observe that models learned incrementally have a more focused attention versus equivalent-architecture counterparts trained simultaneously (as shown in Figure 12). Models with strong peak tuning, which incorporate the surround properly, displays association field effects, focusing on similar patterns as present in the receptor field. Furthermore, because of the incremental freezing scheme between all incremental models, they have the same attention despite variations in the readout layer. However, incremental models display a focused attention, meaning the initially trained SA representation using the CTL (in `rf+sa-CNN*(Incr)`) allows for proper learning of attention weights. This further supports that the 30 to 1 paramter bottleneck in the CTL is not a limiting factor, and the the gain in performance in the latter incremental models (in `ff+sa-CNN*(Incr.FC1)` and `ff+sa-CNN*(Incr.FC2)`) are associated with the FCL.

A.7 SELF-ATTENTION CNNS ARE DATA EFFICIENT

Table 3: Average Pearson correlation and peak tuning metrics for models trained at different data sizes of M1S1. Correlation SEM $= 0.008$ was consistent across models.

	25% of M1S1			50% of M1S1		
Model	Δ CORR. of `ff-CNN`	PT_J	PT_S	ΔCORR. of `ff-CNN`	PT_J	PT_S
`ff-CNN`	0.0%	0.232	1.026	0.0%	0.762	1.556
`ff+sa-CNN`	+7.2%	0.927	1.656	+7.6%	1.026	2.715
`rf-CNN`	+26.9%	0.000	0.000	+21.1%	0.000	0.000
`rf+sa-CNN`	+23.6%	0.000	0.000	+19.8%	0.000	0.000

We trained baseline feedforward CNN models and their counterparts with self-attention at various training dataset sizes. We conclude that percentage improvements over `ff-CNN` are furthered at lower data constraints (as shown in Table 3), alluding to the potential efficiency of SA in accumulating surround information compared to other context mechanisms. We note that `rf-CNN` is the most data efficient when evaluated solely on Pearson correlation. However, it is important to see from Table 3 that at 25% and 50%data, `rf-CNN` and `rf+sa-CNN` completely fail in the peak tuning index, indicating that the models were entirely unable to model the response magnitude of the highest excitatory images. This lends to our claim that although `rf-CNN` does well in correlation, contextual information (as is present in `ff-CNN` and `ff+sa-CNN`) is necessary to capture peak responses.

A.8 POPULATION TUNING CURVES

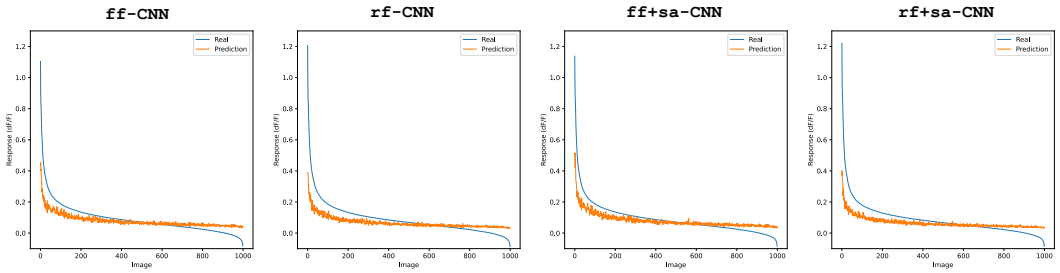


Figure 13: Population average tuning curves for M1S1.

Differences in peak tuning can also be observed in the population tuning curves (see Figure 13). Average curves are derived by calculating rank ordered tuning curves for each neuron individually, then averaging over the image number across neurons.

A.9 V MAPPING PARAMETER γ IN SELF-ATTENTION

In the self-attention layer (as shown in Figure 2), the V mapping parameter γ allows further factorization of inter-channel mixing and spatial interactions. Toggling $\gamma =$ False removes the transformed value vector, and attention weights instead directly on the input representation. Note that the γ parameter does make a difference performance wise. The only difference between `rf+sa-CNN`* and `rf+sa-CNN` is the presence of a SA with γ = True block in the former and a SA with γ = False block in the latter. `rf+sa-CNN`* has better correlation, PT_J, and PT_S values, meaning allowing for the V mapping in SA allows for more flexibility, despite the lack of a 3×3 convolution and FCL layer in these models.

A.10 CTL CHANNEL NUMBER BOTTLENECK

In models with a CTL readout, the final layer is performing a 30 to 1 or 32 to 1 weighted sum, depending on the number of channels. Thus, an issue we considered was that such a narrow final layer would inhibit proper backpropagation of error signals to upstream modules. To address the concern of a $30 \rightarrow 1$ mapping in the CTL layer being too tight of a an initial bottleneck, we trained `rf+sa-CNN` with $c = 375$ channels, so that it would have a $375 \rightarrow 1$ CTL mapping instead. The results were comparable to the `rf+sa-CNN` with $c = 30$, meaning the drop in performance from `rf-CNN` cannot be attributed to a parameter bottleneck.

A.11 IMPORTANCE OF POST-SELF-ATTENTION CHANNEL MIXING

Additionally, we tested self-attention models without post-SA convolutions (i.e. no βCPB layers) and observed sharp drops in performances compared to baseline CNNs. This suggests that inter-channel mixing is crucial in processing the output of self-attention into a interpretable representation by the readout layer. We note that transformer block in modern computer vision models employ a multi-layer perceptron immediately after self-attention, which aligns with our findings.

To compare the importance of the 3×3 versus 1×1 kernel size and FCL vs CTL readout as a means of incorporating surround information, we compared the following models: [αCPB $\rightarrow$ αCPB $\rightarrow$ SA(γ = True) $\rightarrow$ βCPB($k = 1$)$\rightarrow$ βCPB($k = 1$) $\rightarrow$ FCL] $\overset{\text{vs}}{\Longleftrightarrow}$ [αCPB $\rightarrow$ αCPB $\rightarrow$ SA(γ = True) $\rightarrow$ βCPB($k = 3$)$\rightarrow$ βCPB($k = 3$) $\rightarrow$ CTL].

The former with βCPB($k = 1$) and FCL outperformed the latter with βCPB($k = 3$) and CTL. Thus, direct access to all spatial features with a fully connected layer is stronger than convolving the surround into the center. We observe that the FCL is the strongest factor for predicting the peak responses, and is bolstered by the addition SA, as `ff+sa-CNN` outperforms `ff-CNN` in peak tuning.

A.12 RECEPTIVE FIELD CHARACTERISTICS AND POPULATION-LEVEL RF DISTRIBUTION

We mapped the neural networks' receptive fields using optimally oriented short bars. The left and center panel of Figure 14 illustrates two example neurons' receptive fields, indicating that a short bar outside the classical receptive fields does not elicit a response greater than the baseline. The half-height maximum receptive field diameter is approximately 3 pixels (0.4 degree visual angle), corresponding to a 2-STD contour diameter of about 0.8 degree visual angle, consistent with experimentally determined receptive field sizes. The right panel of Figure 14 displays the distribution of mapped receptive fields of neurons for one site.

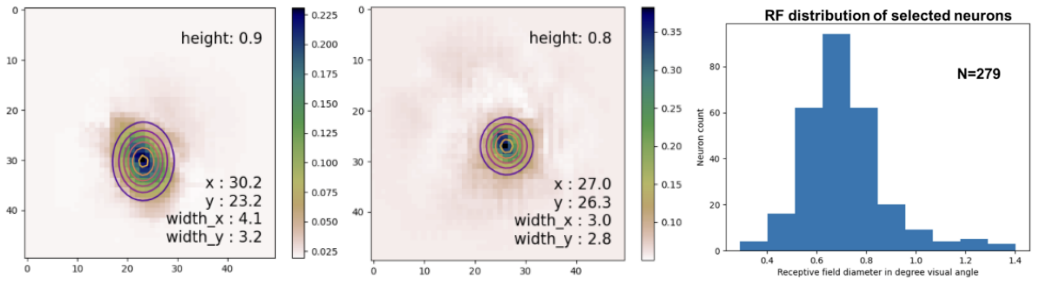


Figure 14: Receptive field properties of neurons and their population-level distribution. **Left and Center:** two sample neurons' response maps fitted with an elliptical Gaussian. **Right:** histogram showing the distribution of receptive field diameters (in degree visual angle) for 279 neurons.

A.13 CAPTURING MODULATORY PROPERTIES OF EXTRA-CLASSICAL RECEPTIVE FIELDS

To verify that our neural networks exhibit the extra-classical receptive field contextual modulation phenomena noted by the reviewer, we replicated following experiment by (Cavanaugh et al., 2002) on our digital neurons.

The following describes the methodology of the experiment. Neurons were presented with center grating stimuli in optimal orientation and spatial frequency of varying diameters (center-only stimuli), centered on their receptive fields. The diameter of the smallest center-only stimulus eliciting at least 95% of the neuron's maximum response defines the GSF (grating summation field). Each neuron was also presented with surround grating stimuli (with gray apertures centered on the receptive field) of varying sizes. The classical receptive field was estimated as the aperture diameter at which the annular stimulus response is significantly below the neuron's maximum response to a circular grating patch. This aperture, referred to as the AMRF, represents where surround-only stimuli do not elicit significant responses above the baseline response. When the center-only stimulus is significantly larger than the size of the classical receptive field, surround suppression—a characteristic example of extra-classical receptive field modulation—is observed.

Figure 15A shows four example neurons from Cavanaugh et al.'s study demonstrating the classical surround suppression phenomena. Figure 15B and 15C shows our replication of their experiment, comparing results for the baseline model (ff-CNN) and the self attention model (ff+sa-CNN). The classical receptive field size of our neurons estimated using AMRF was approximately 0.8 degree visual angle. We found that when the center-only stimulus exceeded this classical receptive field size, surround suppression—a hallmark of extra-classical receptive field modulation—was consistently observed.

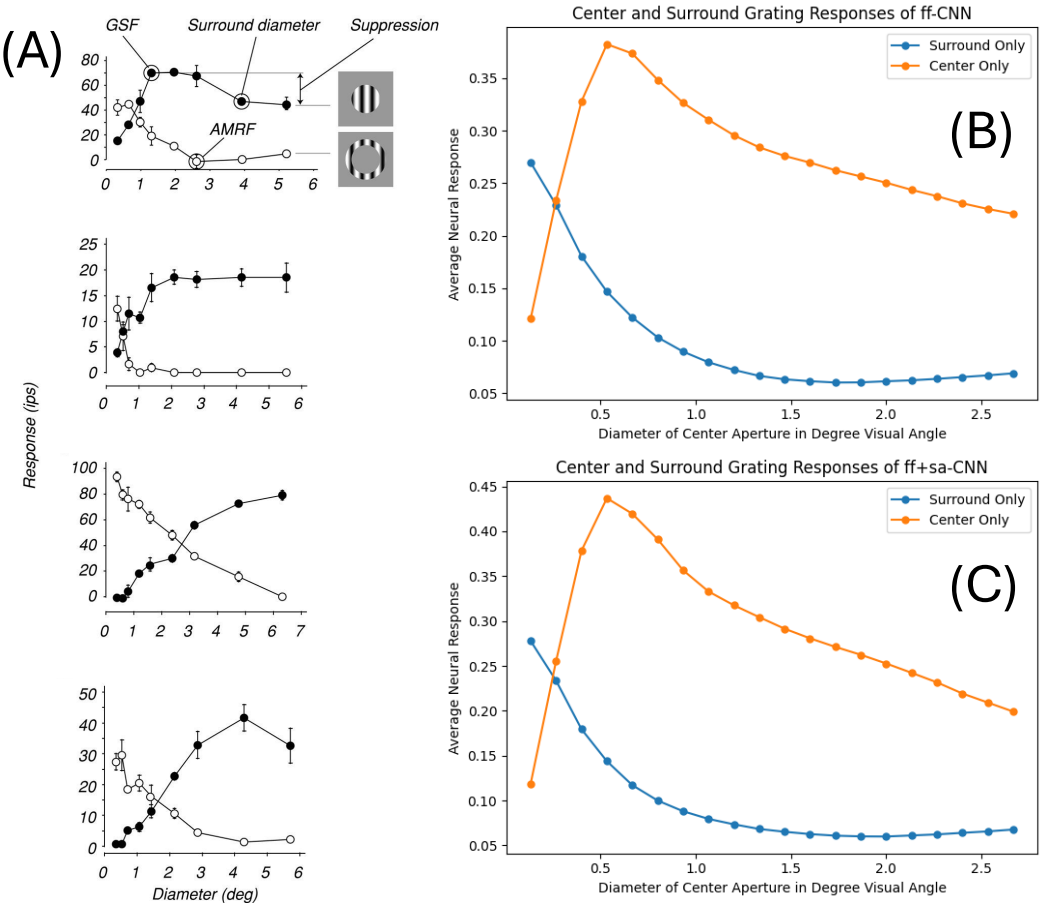


Figure 15: Capturing basic modulatory nature of extra-classical receptive fields. **(A)** shows four example neurons from (Cavanaugh et al., 2002) study demonstrating the classical surround suppression phenomena. **(B), (C)** shows the average population response to the center only stimuli and the surround only stimuli for two models, `ff-CNN` and `ff+sa-CNN`.

A.14 PERFORMANCE COMPARISON OF DIFFERENT MODELS AT VARYING DATA SIZES

Table 1: M2S1 Performance Metrics when Trained with 12.5% Data

Model	CORR	PT_S	PT_J
`ff-CNN`	0.260	0.000	0.000
`ff+sa-CNN`	0.272	0.000	0.000
Recurrent CNN (1 layer, 7 iterations)	0.291	2.174	1.304

(A)

Table 2: M2S1 Performance Metrics when Trained with 25% Data

Model	CORR	PT_S	PT_J
`ff-CNN`	0.295	0.000	0.000
`ff+sa-CNN`	0.317	0.033	0.013
Recurrent CNN (1 layer, 7 iterations)	0.298	2.174	1.304

(B)

Table 3: M2S1 Performance Metrics when Trained with 100% Data

Model	CORR	PT_S	PT_J
`ff-CNN`	0.474	17.333	8.667
`ff+sa-CNN`	0.498	19.667	11.000
Recurrent CNN (1 layer, 7 iterations)	0.435	5.667	3.333

(C)

Figure 16: Performance of `ff-CNN`, `ff+sa-CNN`, and recurrent CNN at varying data sizes. **(A)** shows performance metrics at 12.5% data. **(B)** shows performance metrics at 25% data. **(C)** shows performance metrics at 100% data

Our results indicate that when the full training dataset (100% data) is utilized, the feedforward CNN augmented with self-attention (`ff+SA-CNN`) outperforms the recurrently augmented CNN across both overall tuning metrics and peak tuning metrics. However, consistent with the findings of Zhang et al. (2022), the recurrent augmented CNN exhibits superior data efficiency. Specifically, when the training dataset size is reduced to 25% or less, the recurrent augmented CNN surpasses both `ff-CNN` and the `ff+sa-CNN` in performance.

A.15 CODE FOR EXPERIMENTS

The code is hosted at the github repository: https://github.com/lucanren/sacnn