We thank the reviewers for their valuable suggestions and critiques and provide clarifications for their comments here.

First Review:

• *TECHNICAL ISSUES:* In the ImageNet experiment, we only used 500 samples from each class and did not perform any data augmentation during training or multi-cropping during the test. This was done to provide a standard and balanced dataset and a general CNN architecture, i.e., an impartial setting, to evaluate the exclusive effect of adding SM to the network. To follow the suggestion of the respected reviewer, we repeated the ImageNet experiment with a **standard ResNet-18** (with pre-activation and L_2 regularization) and used the same strategy for incorporating the SM kernel. Although ResNet extensively uses batch normalization (BN), which is expected to potentially reduce the gain offered by SM, our SM-ResNet-18 still showed superior performance over the standard ResNet-18 (Fig. a). We repeated this experiment on a smaller dataset obtained by sampling 100 instances from each class. We observed that SM-ResNet-18 offers about 12% gain in relative accuracy over the baseline. This result indicates that SM-CNN is more effective when learning from datasets with smaller sizes and may lose its gain for very large training datasets.

In our implementations, SM was added to the 1st layer as such modulation is more common in the early visual cortex. However, the paper also reports on the result of an SM-CNN variant with SM added to a later layer with results beating the standard CNN. Finding the optimal structure for the SM and its most effective placement in CNN can be studied in the future. We trained all of the models from scratch to better analyze the sole effect of SM on the training procedure.

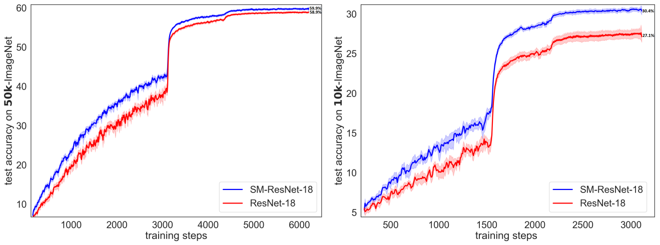
• *BIO-INSPIRED:* As explained in the introduction, surround modulation (SM) occurs when the center and the surround of receptive fields carry similar visual features. Since the units of each feature map in CNNs contain responses to the same features, we implemented SM by incorporating units of each feature map. We were aware that in our linear model, the center might be suppressed even when it is not activated. To account for this, in our first design, the SM unit was implemented as a linear filter followed by a ReLU nonlinearity, which resulted in a more biologically plausible neural activity. However, experiments showed that a linear filter alone offers slightly better accuracy while being simpler.

• *SPARSITY:* One notable aspect of our study is that the SM structure alters neural activities of CNNs and makes them more biologically plausible. The effect of SM on decorrelation and sparsity of neural coding has been widely studied in neuroscience. We have shown that our implementation of SM results in similar outcomes, even though we did not explicitly design it for this purpose. We included the histograms of sparsity and decorrelation to show how our model agrees with those studies (see Gallant et al. 2000, 2002). These diagrams can be summarized in the final version if recommended. Our analysis of sparsity also motivates the question of whether sparsity in neural activity can boost training, another attribute studied in neuroscience (see Yao et al. 2007). As our paper reports, this notion is supported in our study. In other words, the gain in learning speed, especially when using fewer training samples (Fig. a), made our network behave more analogous to the biological systems, even though the method was not designed for such purpose.

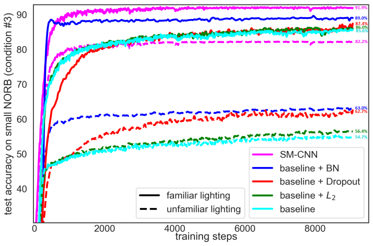
Second Review: The application of the DoG filter in computer vision is in edge detection. We examined this in our experiment and did not find it helpful (see the C_1 variant of SM-CNN in Table 1). We will add a note about this point. In the paper, we provided 2 control models for the baseline and 3 control models for the SM-CNN, to which we will need to add explicit references to Table 1 in the final version. On the issue of control model details, Supp. Materials contains a description about the structure of the networks, from which the effect of each control model on the # of parameters can be easily derived (these effects will be described in the final version). Hyperparameters were roughly tuned for baselines. Our analysis shows that a learning rate of 10^{-4} offers a good tradeoff between speed and robustness for all scenarios. Changing the value of the learning rate does not impact the superiority of our model over the baseline.

• *SUGGESTED CONTROLS:* We designed a new control model based on the suggested **random kernel** matching the 1st and 2nd statistics of the SM kernel, and analyzed it on the ImageNet experiment. The final accuracy was 39.5%, which is lower than both the baseline (43.2%) and SM-CNN (40.9%). We also repeated the experiment on small NORB with 3 control models by adding standard **regularization methods** to the baseline network. In the 1st model, we added Dropout to 3 layers. In the 2nd model, we added BN to all convolution layers. In the 3rd model, we added L_2 regularization to all of the weights. We extended Fig. 4 of the paper by adding these results to it (Fig. b), illustrating that SM-CNN has better generalization in this problem. As mentioned earlier, Fig. a contains the **ResNet** analysis.

Third Review: Thank you for the helpful suggestions. Grating patches with different radius sizes or orientations between surround and center have been widely examined in neurophysiological reports. Performing such analysis and evaluating the similarity of tuning curves with those reports can certainly be targeted as a direction for future work.



(a) ResNet experiments on ImageNet.



(b) Regularization experiments on small NORB.