# Lateral Modulation in Residual Networks: Enhancing Face Recognition Under Occlusion

Heeya Shah & Raiyan Aaijaz
Department of Systems Design University
University of Waterloo
Waterloo, ON

August 4, 2025

**Abstract**

Robust facial recognition under occlusion remains a key challenge in artificial vision systems. We hypothesized that introducing biologically inspired lateral inhibition and excitation mechanisms could enhance convolutional neural networks (CNNs) under such conditions. To test this, we modified ResNet-18 and ResNet-50 architectures by adding a fixed spatial modulation layer that mimics V1 center-surround receptive fields, and trained them on unoccluded images from a subset of the Real-World Occluded Faces (ROF) dataset. We evaluated their performance on neutral, sunglasses, and masked test sets, systematically varying excitation and inhibition strengths. Results showed that lateral modulation improved recognition accuracy under occlusion, especially for masked faces, with optimal parameters differing by model depth and occlusion type. Occlusions of the upper face such as those with sunglasses led to the largest performance drop, highlighting the importance of the eye region in identity recognition. Notably, the models remained stable even under extreme modulation settings. These findings suggest that carefully tuned neurobiological mechanisms can improve CNN robustness and motivate future work on dynamic, learnable modulation layers.

# 1 Introduction

## 1.1 Background and Motivation

Understanding how the human visual system recognizes faces, especially under conditions of partial occlusion, is important both for neuroscience and artificial vision systems. In real-world scenarios, faces are rarely fully visible. Instead, they are often partially obscured by glasses, facial masks, or other

occluding objects, yet humans can often identify individuals with remarkable accuracy. This robustness is attributed to the hierarchical processing of visual information in the brain, where low-level features like edges and orientations are first extracted in the primary visual cortex (V1) and then progressively integrated into complex object representations in higher visual areas [1].

Artificial vision systems, particularly convolutional neural networks (CNNs), are modeled after this hierarchy. Early convolutional layers behave similarly to simple and complex cells in the V1, detecting localized patterns such as edges and textures, while deeper layers combine these features into higher-level representations used for classification [1]. However, unlike the biological system, these networks often experience significantly degraded performance when key features are missing or occluded.

## 1.2 Literature Review

Current research supports the use of standard CNNs for facial recognition tasks under no occlusion conditions as they share a similar hierarchical structure to biological systems [1]. However, the drop in performance under occlusion conditions has been shown to be due to the missing critical features required for holistic pattern recognition [2]. In contrast, the human visual system has developed and evolved mechanisms to mitigate such degradation. Lateral inhibition, a well-established process in V1, helps sharpen feature boundaries by suppressing less relevant signals, reducing noise, and enhancing contrast [1]. Lateral excitation, meanwhile, supports pattern completion by reinforcing co-activation across neighboring spatial regions, helping the brain "fill in the blanks" when only partial features are available [1].

These mechanisms improve robustness in perception, particularly in the case of incomplete or ambiguous inputs. Studies have also shown that the balance of inhibition and excitation is dynamically tuned based on visual context, allowing flexible adaptation [3]. Previous research has explored CNN variants with attention or recurrent feedback, but the explicit simulation of early lateral interactions (especially using biologically inspired tuning) remains relatively underexplored.

## 1.3 Experimental Approach

This project explores whether introducing biologically inspired lateral inhibition and excitation layers into a ResNet-based CNN can improve robustness under partial occlusion. A subset of the ROF dataset, containing 20 celebrities under three occlusion conditions (neutral, sunglasses, and masks) will be used to simulate increasing levels of visual occlusion. The model will be trained only on unoccluded faces and tested on all three conditions to assess generalization under occlusion.

The inhibition and excitation strengths will be systematically varied across a biologically plausible range of -2.0 to +2.0, as well as extreme, unrealistic values of -1000 and +1000, to evaluate both effectiveness and failure modes. Recognition accuracy will be measured across all configurations to evaluate how different lateral balances affect model performance. Through this, we aim to determine which types of occlusion benefit most from lateral modulation, and what that reveals about both artificial and biological visual systems.

This approach contributes to a growing body of work on biologically grounded enhancements to neural networks, with the goal of making CNNs not just more accurate, but more robust to real-world conditions that mimic the human visual system.

# 2 Methods

## 2.1 Model & Baseline

This project utilizes convolutional neural networks (CNNs) for the task of face recognition due to their biological plausibility and proven success in visual processing tasks. CNNs are particularly effective for image classification due to their ability to extract hierarchical spatial features and their parameter-sharing structure, which makes them scalable to high-dimensional inputs [4].In this project, a specific type of CNN known as a residual network (ResNet) was used; ResNets extend standard CNNs by introducing skip connections that enable the training of much deeper models without suffering from optimization instability.

The first baseline model used was ResNet-18, an 18-layer network that served as a control for performance under non-occluded face recognition training and testing, as well as for evaluation under increasing degrees of occlusion. The model was initialized with weights pretrained on ImageNet to facilitate convergence and leverage feature representations learned from large-scale natural image datasets. ResNet-18 is well-suited as a baseline due to its balance of depth and computational efficiency. It was shown to converge faster than its plain counterpart while achieving similar or better accuracy [5]. This highlights its effectiveness as a simple but representative benchmark model, as it allows for an assessment of generalization without architectural complexity. The unmodified ResNet-18 also serves as a control condition, allowing for isolation of performance differences attributable to the lateral intervention.

The second baseline model used was ResNet-50, a much deeper and more expressive network than ResNet-18 [5]. It incorporates additional layers and bottleneck blocks to enhance feature abstraction while maintaining computational efficiency. As with the other baseline model, pretrained weights were used to initialize the network. ResNet-50 was selected to evaluate whether increased depth and representational capacity, in combination with lateral excitation

and inhibition mechanisms, could improve recognition under occlusion. This is supported by [5], which shows that deeper residual networks achieve higher accuracy and more stable optimization without suffering from the degradation issues observed in plain architectures. These properties made ResNet-50 a suitable foundation for testing biologically inspired modifications, particularly in scenarios where parts of the input are missing or degraded. Comparing ResNet-50 against ResNet-18 allowed for the isolation of the effects of both network depth and structural interventions in learning from occluded data.

## 2.2 Network Architecture

CNNs are composed of layered operations designed to extract increasingly abstract visual features from image data. Convolutional layers apply learned filters to detect local patterns such as edges, orientations, and textures. Pooling layers downsample spatial dimensions to reduce computational load and support translation invariance, while fully connected layers, typically near the output, integrate spatial information to perform classification. As network depth increases, the model can capture more complex feature hierarchies [4].

ResNet architectures build on this framework by introducing skip connections that bypass one or more layers. These connections allow the model to learn residual functions, where each block's output is formulated as $F(x) = H(x) - x$, rather than learning the full transformation $H(x)$ directly. This residual formulation improves gradient flow during backpropagation and helps prevent vanishing gradients in deeper networks [5].

ResNet-18 consists of four sequential stages of residual blocks, each containing two 33 convolutional layers and a skip connection. These blocks preserve spatial resolution within each stage. Identity mappings are used when input and output dimensions match, while projection shortcuts are used when they differ.ResNet-50 replaces the basic blocks with bottleneck blocks, which use a 11 convolution to reduce channel dimensionality, a 33 convolution for spatial filtering, and a final 11 convolution to restore the original depth. This design increases depth without a large increase in computational cost and the architecture enables stable training and performs well across standard image classification tasks [5].

These baseline architectures provide the foundation for introducing biologically inspired lateral mechanisms intended to modulate early visual processing. By incorporating these components into the initial stages of both ResNet-18 and ResNet-50, the models were adapted to examine how structured excitation and inhibition influence feature extraction under varying levels of occlusion.

## 2.3 Model Intervention

To enable this investigation, a lateral modulation layer was introduced at the input stage of each network, incorporating spatially structured excitation and inhibition into the processing pipeline. This layer operates independently of the original ResNet modules and applies the same structure across both model variants, ensuring consistency in how lateral interactions influence feature encoding.

The modulation layer is implemented as a fixed 3×3 convolutional filter applied channel-wise, simulating a biologically inspired center-surround receptive field. The kernel consists of a central excitatory weight surrounded by uniform inhibitory weights, reflecting the spatial organization observed in early visual processing. To isolate the effect of spatial structure, the kernel parameters are non-learnable during training. A ReLU activation follows the convolution to suppress negative values, aligning with the non-negative firing rates of biological neurons. The resulting signal can either replace or be added to the original input via a configurable skip connection, providing flexibility in how modulation integrates with the base network.

Excitation and inhibition strengths are parameterized and systematically varied across training runs. The excitation values used were [0.8, 1.0, 1.2, 2.0], and the inhibition values were [-0.1, -0.3, -0.5, -2.0]. These settings include biologically plausible ranges as well as exaggerated values intended to probe the model's sensitivity to different balances. By applying the modulation layer identically in both ResNet-18 and ResNet-50, the intervention design supports controlled comparison across architectures of different depth. This configuration enables targeted evaluation of how spatially structured excitation and inhibition influence early feature integration and downstream classification performance under occlusion.

To further test the model's tolerance to extreme input modulation, additional experiments were conducted using ResNet-50 with non-biological parameter extremes. Specifically, inhibition was set to –1000 with excitation at 0, and vice versa. These configurations were used to evaluate whether excessive suppression or excitation would degrade performance under occlusion. The results from these extreme cases were compared to standard settings to assess the model's stability under exaggerated conditions.

## 2.4 Performance Metric

Model performance was evaluated using classification accuracy, defined as the proportion of correctly classified test images out of the total number of test samples. Accuracy was computed separately for each combination of occlusion condition, model architecture, and inhibition/excitation setting.

For baseline models (ResNet-18 and ResNet-50), both training and testing

accuracy were tracked over the full 10 epochs of training. Accuracy trends were visualized to assess convergence behavior and potential overfitting, with final performance comparisons based on testing accuracy at the 10th epoch. This approach established a reference point under non-occluded and occluded conditions.

For experiments involving the lateral inhibition and excitation intervention, training and testing accuracy were recorded across 10 epochs for each unique combination of excitation and inhibition values. The best testing accuracy from these 10 epochs, usually at epoch 10, was selected as the representative value. Each model trained under a given parameter setting on the neutral dataset was subsequently tested on the sunglasses and masked occlusion sets using the same weights. This process enabled direct evaluation of how each modulation setting generalized across occlusion conditions. Final results were visualized as heatmaps to highlight the relationship between excitation/inhibition strength and model performance across occlusion types. No additional metrics such as precision, recall, or F1-score were included, as the dataset was balanced across classes and the evaluation was centered on overall recognition accuracy rather than per-class variation.

## 2.5 Dataset

This project utilized the Real-World Occluded Faces (ROF) dataset, which contains labeled images of celebrities under varying real-world occlusion conditions, including unoccluded (neutral), sunglasses, and masked faces [6]. The original dataset includes images for 180 identities, with 70 of those having both types of occlusion image samples.

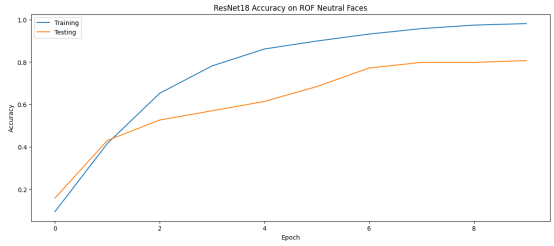
To reduce computational demands and enables extensive experimentation under time and GPU constraints, a subset of 20 identities was selected. These individuals were chosen based on having the highest number of high-quality, neutral images and at least 2 of each occlusion image samples. This ensured a consistent and balanced sample across conditions. Each image was manually reviewed, and entries that did not match the selected identity or included eros such as extreme blur and/or incorrect labels were discarded.

The filtered dataset was then divided into four subsets: neutral_train, neutral_test, sunglasses, and masked. For the neutral condition, an 80/20 split was used for training and testing within each identity class. The sunglasses and masked subsets were used exclusively for testing and were held out during training to simulate real-world generalization to occluded inputs. On average, each identity included approximately 20 training images and 5 test images for the neutral condition, and more than 2 test images for each occlusion variant. While the final dataset was relatively small, it was sufficient for controlled experimentation and parameter sweeps across multiple network configurations.

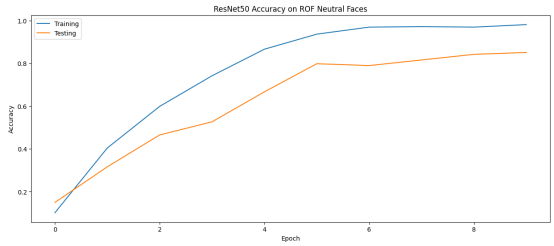
# 3 Results

## 3.1 Baseline Models

As a baseline, ResNet-18 and ResNet-50 were selected for comparison. The final classification layers of model models were modified to match the number of identity classes in the dataset. These baseline models were first evaluated on unoccluded facial images to establish their recognition performance under ideal conditions. Figures 1a and 1b display the training and testing accuracy over 10 epochs for ResNet-18 and ResNet-50, respectively, using the neutral subset of the dataset.




(a) ResNet-18 performance




(b) ResNet-50 performance

Figure 1: Training and testing accuracies of baseline models over 10 epochs on ROF Neutral Face Dataset

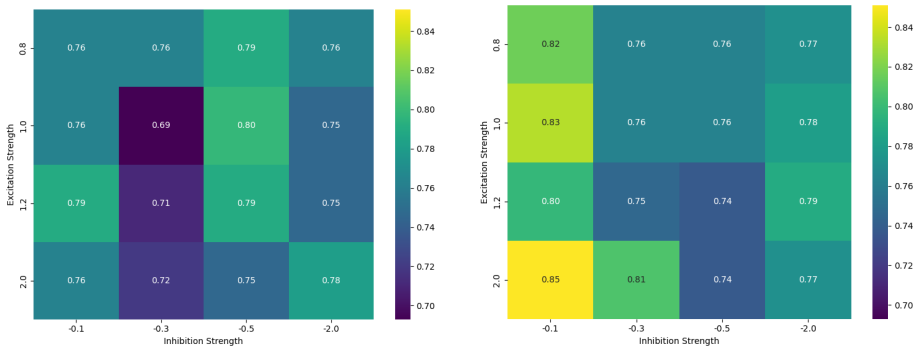
Training was conducted over 10 epochs for both ResNet-18 and ResNet-50. This duration provided sufficient convergence for both models, with stable training and testing accuracy and no signs of severe overfitting. All subsequent experiments were also run for 10 epochs to maintain consistency. The testing accuracy scores for each of the three occlusion conditions (neutral, sunglasses, and masked) using the baseline ResNet-18 and ResNet-50 models are summarized in Table 1.

Table 1: Recognition Accuracy Across Occlusion Conditions for Baseline ResNet Models

| Model | Training Accuracy | Testing Accuracy Score | | |
|---|---|---|---|---|
| | No Occlusion | Neutral Faces | Sunglasses | Masked |
| ResNet-18 | 0.981 | 0.807 | 0.283 | 0.350 |
| ResNet-50 | 0.981 | 0.851 | 0.277 | 0.425 |

## 3.2 Lateral Inhibition Model

A lateral inhibition layer was incorporated into the modified "Inhibited ResNet" architecture, using both ResNet-18 and ResNet-50 as baselines. Each model was trained for 10 epochs, and a grid of inhibition and excitation strength values was tested to evaluate their effect on recognition performance. For each occlusion condition, all combinations of inhibition and excitation were evaluated, and the best testing accuracy observed across the 10 epochs (which consistently occurred at epoch 10) was recorded. The resulting performance is visualized as heatmaps in Figure 2a and 2b, illustrating how different levels of inhibition and excitation affected model accuracy under each condition.




(a) ResNet-18 accuracy

(b) ResNet-50 accuracy

Figure 2: Best test accuracies at epoch 10 for ResNet-18 and ResNet-50 across inhibition and excitation strength combinations under the neutral condition

For the neutral condition, the highest testing accuracy for the ResNet-18 model was 0.80, achieved with excitation at 1.0 and inhibition at –0.5. The ResNet-50 model reached a maximum accuracy of 0.85 using the maximum excitation of 2.0 and minimal inhibition of –0.1.

The Inhibited ResNet models were also evaluated across all combinations of inhibition and excitation for the sunglasses and masked conditions. The corresponding heatmaps are provided in Figures 1 and 2 in the Appendix, and

the summarized accuracy results are shown in Table 2.

Table 2: Best Accuracy and Associated Inhibition/Excitation Values by Model and Occlusion Type

| Occlusion Conditions | ResNet-18 | | | ResNet-50 | | |
|---|---|---|---|---|---|---|
| | Excitation | Inhibition | Accuracy | Excitation | Inhibition | Accuracy |
| Sunglasses | 0.8 | -0.1 | 0.31 | 2.0 | -0.3 | 0.39 |
| Masked | 1.0 | -2.0 | 0.52 | 1.2 | -0.1 | 0.54 |

### 3.3 Extreme Parameter Testing

To further examine the model's behavior, a version of the Inhibited ResNet-50 was evaluated using extreme values of inhibition and excitation. This model was trained on the neutral dataset and tested across the neutral, sunglasses, and masked conditions to assess whether excessively large parameter values would affect performance. Specifically, two configurations were tested: inhibition set to –1000 with excitation set to 0, and excitation set to 1000 with inhibition set to 0. These tests were conducted only on the ResNet-50 model. The highest testing accuracy for each occlusion condition using these extreme parameters is reported in Table 3.

Table 3: Testing Accuracy of Inhibited ResNet-50 with Extreme Inhibition/Excitation Parameters

| Model Parameters | | Occlusion Conditions | | |
|---|---|---|---|---|
| Excitation | Inhibition | Neutral | Sunglasses | Masked |
| 0 | 1000 | 0.780 | 0.307 | 0.457 |
| 1000 | 0 | 0.817 | 0.314 | 0.448 |

## 4 Discussion

### 4.1 GPU Constraints and Experiment Limitations

Before analyzing the results, it is important to acknowledge several limitations that impacted the experiment. A primary constraint was the small dataset size used for both training and testing. From the ROF dataset, a subset of 20 celebrities with the highest number of neutral images was selected. The data was manually cleaned and categorized into four subsets: neutral_train, neutral_test, sunglasses, and masked. An 80/20 train-test split was applied to the neutral images for each identity to maintain balance, resulting in approximately 20

training images and 5 test images per celebrity across each occlusion condition. While sufficient for experimental comparison, this remains a small dataset for training deep facial recognition models.

The size of the dataset was reduced primarily due to GPU limitations. Preliminary tests on the full set of 60 celebrities resulted in training times exceeding two hours per model, making it impractical within the time constraints of the project. By reducing the dataset to 20 of the most represented identities, the training time per model decreased significantly to approximately 30 minutes per run, enabling a feasible workflow for testing multiple combinations of inhibition and excitation parameters. This trade-off was necessary to explore the lateral modulation effects efficiently across models and occlusion conditions.

However, this limitation also affects the generalizability of the results. In particular, the inhibited ResNet models were only evaluated using a single training run per parameter configuration. Ideally, each combination of excitation and inhibition should have been trained across multiple trials, with the mean and standard deviation of test accuracy reported to ensure statistical reliability. Due to resource constraints, this was not feasible. Each model took approximately 20-30 minutes to train on the available GPU setup, and with 4 excitation and 4 inhibition values, a total of 16 unique configurations required evaluation. Running five trials per configuration was too time consuming for the scope of this project. As a result, only one trial per configuration was conducted, and the best epoch accuracy was reported. Future work should include repeated runs with statistical averages to strengthen the validity of the findings.

## 4.2 Baseline Model Performance

Both ResNet-18 and ResNet-50 baseline models were trained for 10 epochs, achieving a maximum training accuracy of 0.981 on the final epoch. Since both models reached the same training accuracy, this allowed for a fair comparison of their general performance despite their architectural differences. Accuracy curves stabilized near the 10th epoch, with earlier epochs showing gradual improvements. This confirmed that 10 epochs was an appropriate training duration as it was sufficient for convergence while remaining computationally feasible.

Each model was evaluated under all three occlusion conditions, as shown in Table 1. ResNet-18 achieved a testing accuracy of 0.807 on unoccluded faces, 0.283 on faces with sunglasses, and 0.350 on faces wearing masks. In comparison, ResNet-50 achieved 0.851 on neutral images, 0.277 on the sunglasses condition, and 0.425 on the masked condition.

In two out of the three conditions, ResNet-50 outperformed ResNet-18, most notably on the masked dataset, where it achieved a 7.5% higher accuracy. This improvement is likely due to ResNet-50's increased depth and its ability to extract richer mid and high-level features, which can help reconstruct partially occluded facial patterns. Similarly, in the neutral condition, ResNet-50 showed a moderate

advantage, which can be attributed to its greater capacity for fine-grained feature extraction when full facial information is available.

In the sunglasses condition, however, ResNet-18 slightly performed better than ResNet-50, with a 0.6% higher accuracy. Given the small size of the dataset and the minimal difference in performance, this result is likely due to random variation, such as data shuffling or batch-level effects, rather than a meaningful architectural advantage. Overall, these results are consistent with prior research: models with deeper layers tend to perform better in more visually complex scenarios, particularly when critical facial features are partially occluded [7].

### 4.3 Inhibited ResNet - Neutral Condition

By incorporating an inhibition layer into the standard ResNet architecture, positioned directly after the input, clear improvements were made under occlusion conditions (discussed in section 4.4). However, it did not lead to significant gains on neutral images across all inhibition and excitation combinations. Instead, improvements were highly dependent on the specific tuning of excitation and inhibition, and the optimal values varied not only by occlusion condition, but also between ResNet-18 and ResNet-50.

For the neutral dataset, the best accuracy for each model was consistently reached at the 10th epoch. As shown in Figure 2a, the highest testing accuracy for ResNet-18 was 0.80, achieved with an inhibition of –0.5 and excitation of 1.0. In contrast, ResNet-50 achieved its peak performance of 0.85 with minimal inhibition (–0.1) and maximum excitation (2.0).

The difference in optimal parameter values between the two models likely stems from how each architecture processes and integrates features. ResNet-18, with its shallower depth, benefits more from moderate inhibition, which helps sharpen feature boundaries and reduce ambiguity in early layers. Excessive excitation likely introduced noise or blurred feature maps that it could not adequately refine. Conversely, ResNet-50, due to its greater depth, can rely on deeper layers to reconstruct higher-level patterns and integrate spatially distant cues. Thus, it benefits more from strong lateral excitation, which reinforces co-activated spatial features and supports feature grouping and pattern continuity, while only requiring mild inhibition to suppress redundant activations [8].

Since the modified model achieved high performance under the neutral condition, both versions of the Inhibited ResNet were further evaluated on the sunglasses and masked occlusion test sets. Heatmaps were generated to visualize the performance across all inhibition and excitation combinations for each model and condition.

### 4.4 Inhibited ResNet - Occluded Conditions

The Inhibited ResNet models demonstrated improved performance under occlusion compared to their respective baseline versions without lateral inhibition and excitation. As shown in Table 2, both ResNet-18 and ResNet-50 achieved higher accuracy across the sunglasses and masked conditions when the inhibition layer was incorporated. However, the optimal excitation and inhibition values differed depending on both the model and the occlusion type.

In the sunglasses condition, ResNet-18 achieved its highest accuracy of 0.31 using excitation at 0.8 and inhibition at -0.1, showing an improvement from its baseline score of 0.283. This result contrasts with the neutral condition, where stronger inhibition (-0.5) was optimal. The reduction in inhibition under occlusion likely allowed more relevant features to pass through, helping the shallower ResNet-18 retain important facial cues despite partial obstruction. ResNet-50, by contrast, achieved an accuracy of 0.39 in the sunglasses condition, but required maximum excitation (2.0) and increased inhibition (-0.3) compared to the neutral condition, where inhibition was -0.1. The sunglasses occlude the eyes, a critical facial region, so the additional inhibition likely helped suppress misleading signals from the edges of the sunglasses, while the model's deeper architecture preserved enough remaining features to maintain recognition performance.

For the masked condition, the results were even more pronounced. ResNet-18 achieved its best accuracy of 0.52 with strong inhibition (–2.0) and moderate excitation (1.0). The high inhibition helped filter out ambiguous or irrelevant information from the occluded lower face including the celebrity's mouth, jawline and nose, allowing the model to focus on the most discriminative visible features, such as the eyes and eyebrows. The moderate excitation reinforced nearby features without introducing excessive noise. ResNet-50 reached a top accuracy of 0.54 with low inhibition (–0.1) and slightly elevated excitation (1.2). This configuration allowed the deeper model to preserve more spatial information and reconstruct missing facial patterns through its higher-level abstractions. The slight boost in excitation was sufficient to support pattern completion, while the mild inhibition avoided excessive suppression of partially visible inputs.

When comparing both occluded conditions to the neutral baseline, accuracy improved with lateral modulation, but neither model was able to improve the performance so much so that it matched the unoccluded dataset. This indicates that while the addition of inhibition and excitation improves recognition under occlusion, some information loss due to occlusion remains unavoidable. Additionally, performance in the sunglasses condition remained significantly lower than in the masked condition, suggesting that the upper half of the face, particularly the eyes and surrounding features, plays a more critical role in facial recognition than the lower half of one's face.

Finally, the optimal inhibition and excitation values varied by model and condition, highlighting key architectural and biological differences. ResNet-

18, lacking deeper abstraction capabilities, relied on stronger inhibition to sharpen edges and suppress noise, while needing moderate excitation that avoids overactivation. ResNet-50, with its deeper structure, generally required less inhibition but benefited from higher excitation to enhance partially visible patterns. These trends emphasize that tuning lateral interactions can be model architecture and context-specific, and that biologically inspired mechanisms can effectively enhance robustness in artificial vision systems when carefully adapted.

## 4.5 Literature Comparison

The results observed in this study show strong parallels to known mechanisms in the biological visual system and in how the brain processes faces under occlusion conditions. In the human visual cortex, lateral inhibition and lateral excitation play essential roles in enhancing perception by sharpening feature boundaries, reducing noise, and filling in missing information, all essential in the V1 where edge detection and contrast enhancement are crucial for facial recognition [9].

This was reflected in ResNet-18, which consistently benefited from stronger inhibition, particularly under occlusion. This mirrors the function of inhibitory interneurons in V1, where early-stage neurons rely on local contrast to extract useful visual information [1]. Conversely, ResNet-50, with its deeper architecture, benefited more from lateral excitation, which reinforces co-activated features across space, aiding in pattern completion and feature grouping. This mirrors processing in higher visual areas such as V2 and the inferotemporal cortex, where neurons integrate global features and compensate for occluded or missing information [1].

The biggest takeaway from this research is the observation that occluding the eyes such as with the sunglasses condition caused greater performance degradation than occluding the lower face. This result is consistent with human studies showing that the eyes and upper facial region carry more identity-specific information [10]. This reinforces the idea that not all occlusions are equal, regardless of the size of the occlusion. Instead, recognition systems, both biological and artificial, rely more heavily on certain facial features than others for facial recognition.

Lastly, the fact that optimal excitation and inhibition values varied by condition and the model type supports findings that lateral interactions are dynamically adjusted based on context, rather than fixed [11]. This context-specific adaptation not only enhances robustness and efficiency in human vision but also appears beneficial in artificial models as well.

## 4.6 Model Boundary Testing

To evaluate the model's robustness, extreme values of inhibition and excitation were tested independently to observe their effect on performance under occlusion. As displayed in Table 3, although there was a slight drop in testing accuracy, no significant degradation occurred when the model was trained and tested under these extreme conditions.

Under extreme inhibition, one would expect almost total suppression of early feature maps, potentially reducing activations to zero and compromising performance. Likewise, extreme excitation should cause overactivation, leading to a loss of contrast and saturation in the feature maps. However, neither scenario resulted in a substantial drop in accuracy.

This may be attributed to the internal regulatory mechanisms within the network. Layers such as ReLU clip negative activations, Batch Normalization re-centers and scales outputs, and residual connections allow clean signal paths to bypass distorted layers []. Additionally, the optimizer and weight decay may prevent long-term harm from extreme spikes in activation [12] [13]. Another factor may be the clarity and simplicity of the dataset itself. Since the images in the dataset are well-lit and relatively clean, even distorted early features may retain enough useful signal for accurate classification. It is likely that testing on more realistic or challenging datasets, with variations in lighting, blur, or noise, would expose greater sensitivity to extreme parameter values and lead to more pronounced performance degradation.

## 4.7 Future Work

While this project demonstrated that incorporating biologically inspired lateral inhibition and excitation can improve convolutional neural network performance under occlusion conditions, several areas remain for future exploration. First, using a larger and more diverse dataset, with multiple trials would allow for more robust evaluation and better generalization. This could include variations in lighting, pose, blur, and background to simulate real-world conditions more closely. Second, incorporating learnable inhibition and excitation parameters, rather than fixed scalar values, could enable the network to dynamically adapt its lateral modulation during training, making it even more biologically inspired. Additionally, further investigation into layer-wise placement of lateral interactions including earlier layers but not the very first one, may reveal more effective strategies for enhancing robustness. Finally, comparing this biologically motivated approach to other occlusion-handling techniques, such as attention mechanisms or inpainting models, would help assess its practical value within the broader field of occlusion-resilient vision systems.

# References


[1] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth, *Principles of Neural Science, Fifth Edition*. McGraw Hill Professional, 2013.

[2] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," *arXiv preprint arXiv:1908.06290*, 2019.

[3] A. Angelucci and J. Bullier, "Reaching beyond the classical receptive field of v1 neurons: horizontal or feedback axons?" *Journal of Physiology-Paris*, vol. 97, no. 2–3, pp. 141–154, Mar. 2003.

[4] K. O'Shea and R. Nash. (2015) An introduction to convolutional neural networks. arXiv. Accessed: 2025-04-22. [Online]. Available: https://arxiv.org/abs/1511.08458

[5] K. He, X. Zhang, S. Ren, and J. Sun. (2015) An introduction to convolutional neural networks. arXiv. Accessed: 2025-04-22. [Online]. Available: https://arxiv.org/abs/1512.03385

[6] M. E. Erakin. (2021) Real world occluded faces (rof). Accessed: 2025-04-22. [Online]. Available: https://github.com/ekremerakin/RealWorldOccludedFaces

[7] R. Rothe. (2025, Jan.) The magic of hidden layers in neural networks - demistify - medium. Accessed: 2025-04-22. [Online]. Available: https://medium.com/demistify/the-magic-of-hidden-layers-in-neural-networks-989b05791dc7

[8] Y. Chen, S. Liu, D. Zhao, and W. Ji, "Occlusion facial expression recognition based on feature fusion residual attention network," *Frontiers in Neurorobotics*, vol. 17, Aug. 2023.

[9] H. K. Hartline and F. Ratliff, "Inhibitory interaction of receptive fields of the retina," *Journal of the Optical Society of America*, vol. 47, no. 6, pp. 583–594, 1957.

[10] B. P. Quinn and H. Wiese, "The role of the eye region for familiar face recognition: Evidence from spatial low-pass filtering and contrast negation," *Quarterly Journal of Experimental Psychology*, vol. 76, no. 2, pp. 338–349, Feb. 2022.

[11] A. Angelucci and P. 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," pp. 93–120, Jan. 2006.

[12] J. Bjorck, C. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," *Advances in Neural Information Processing Systems*, vol. 31, 2018. [Online]. Available: https://papers.neurips.cc/paper/7996-understanding-batch-normalization.pdf

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

# Appendix

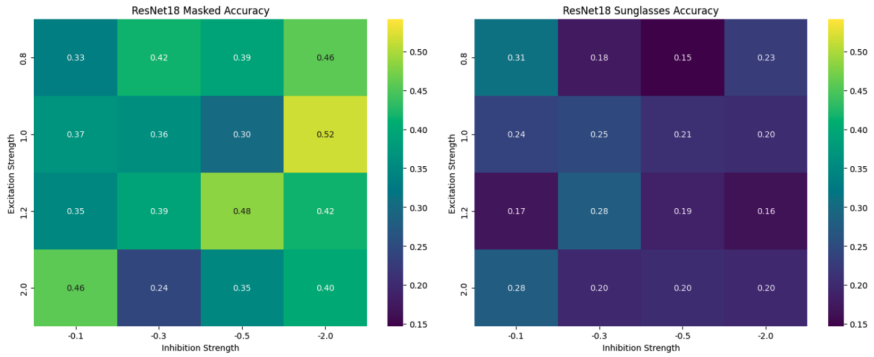



Figure 1: ResNet-18 Masked and Sunglasses Accuracy Heatmap

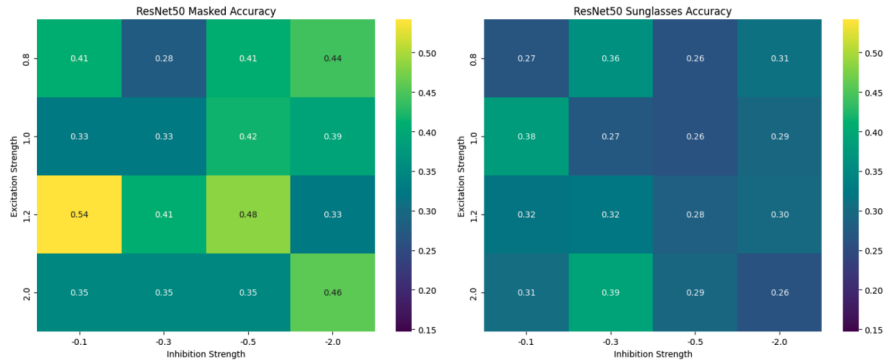



Figure 2: ResNet-50 Masked and Sunglasses Accuracy Heatmap