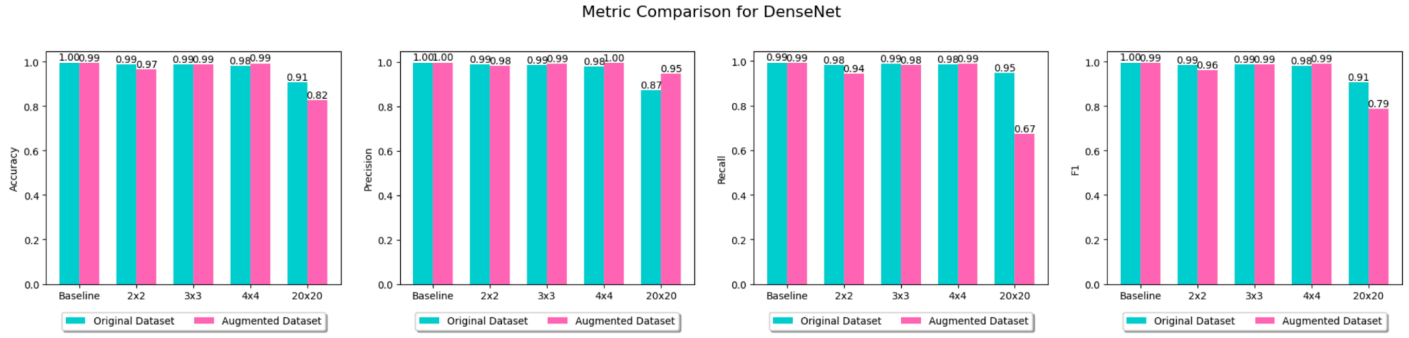


In summary, augmentation techniques notably enhance GoogleNet's classification performance in handling complex permutations, such as the 20x20 tiles, reinforcing the model's adaptability to varied spatial configurations. However, DenseNet-121 demonstrates a slight decline with augmented data for certain permutations, suggesting an optimal balance of complexity and generalization already exists within its architecture. These findings highlight that augmentation is not a one-size-fits-all solution, and its effectiveness is highly dependent on the architecture and the complexity of the data it is applied to.

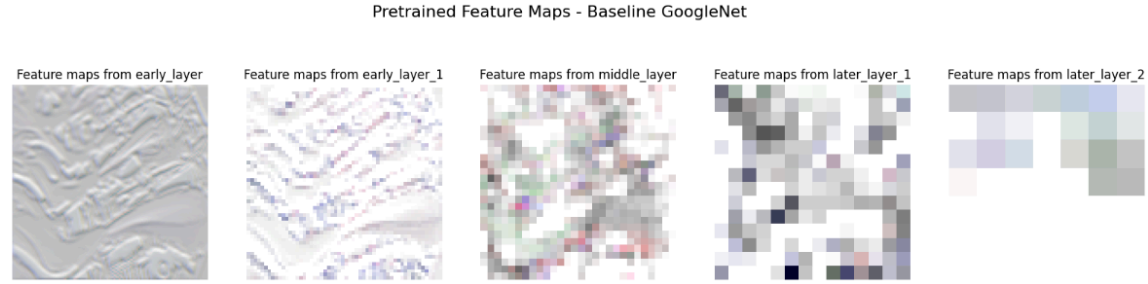
4.3 Feature and Saliency Maps

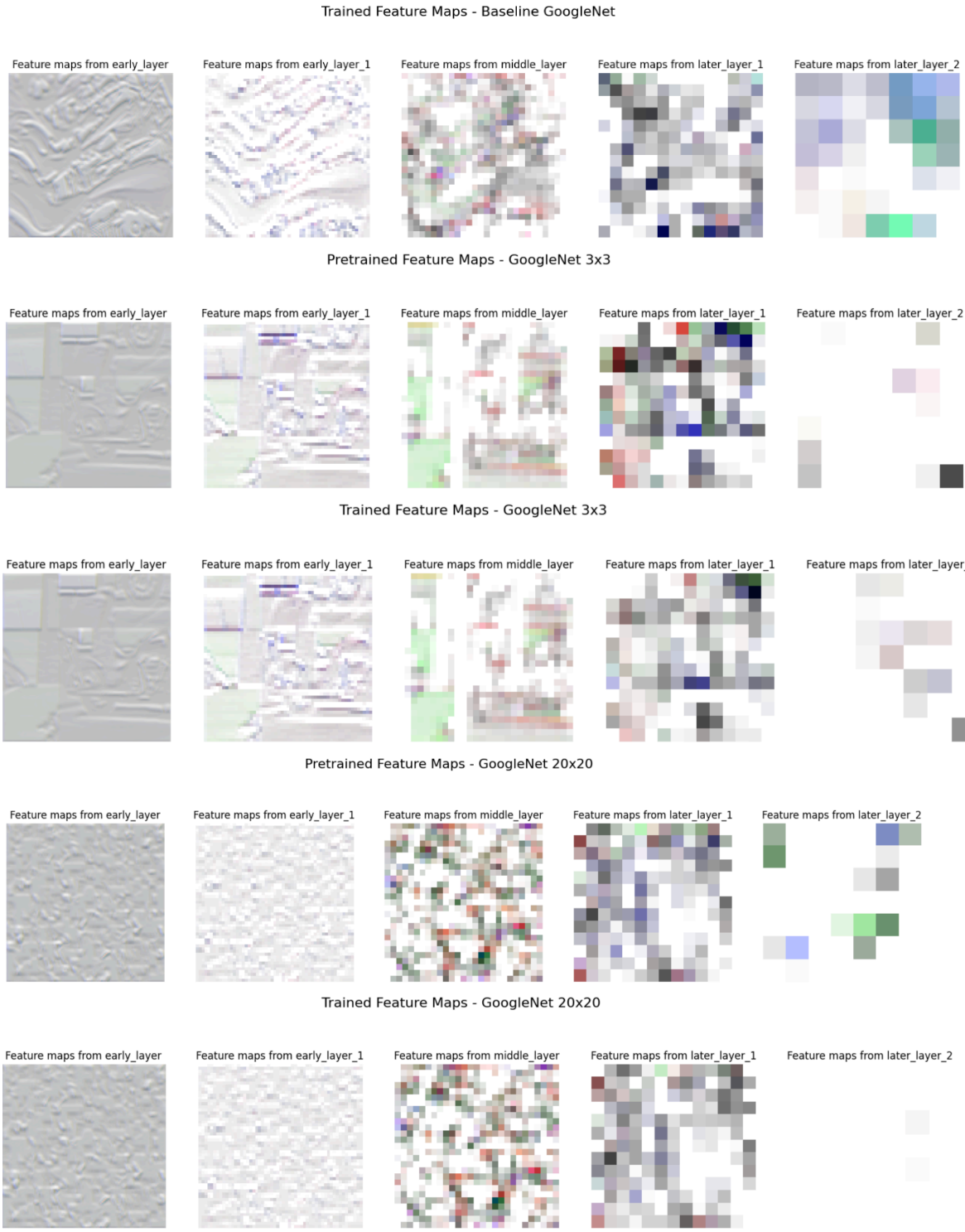
In this section, we'll examine the feature and saliency maps at various tile resolutions to dissect how the convolutional neural networks process these permutations. We're comparing models pretrained to those further honed on our specific data. Our goal is to understand how tile resolution impacts the network's interpretability and learning.

*Due to the size of each plot and the similarity in conclusions across all resolutions, this section will only display plots for baseline, 3x3, and 20x20 resolutions. Plots for other resolutions are available in the code notebook.

4.3.1 Feature Maps

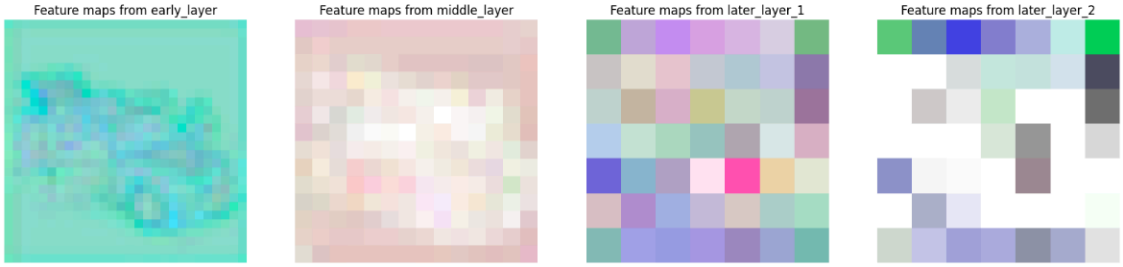
We anticipated that the GoogleNet pre-trained model would manifest broad and less specialized feature activations, not finely attuned to the specifics of the transformed tile datasets. However, in practice, it appears that the layers remain consistent across all permutations, with notable distinctions observed primarily in the deeper layers. This observation underscores that lower layers predominantly capture simpler and more general features, such as edges and colors, while higher layers, which develop abstract representations, exhibit more significant transformations when the network is trained on specific data. Moreover, the tile size exerts impact on the feature maps- Baseline tiles tend to elicit more uniform and less varied feature activation, representing standard, unaltered images. In contrast, 3x3 tiles, and particularly 20x20 tiles, introduce greater complexity due to permutation, resulting in feature maps characterized by enhanced diversity and specificity. Notably, in 20x20 tiles, even in the early layers, it becomes challenging to discern identifiable shapes, unlike the clearer shapes observed in 3x3 and baseline tiles.



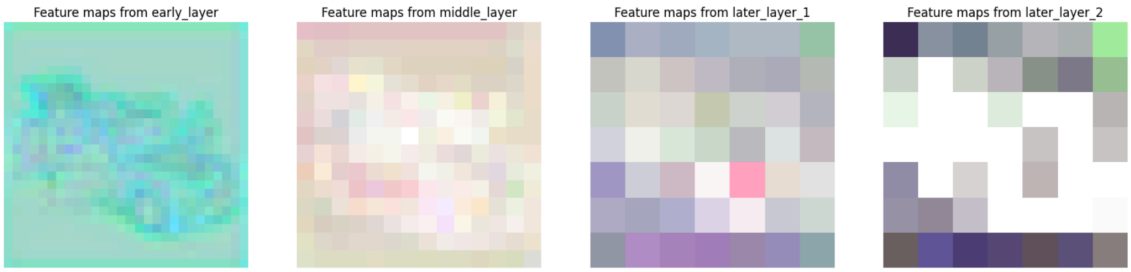


Similarly, for DenseNet, the layers appear consistent across all permutations, except for the deeper layers. Also, there is a discernible difference in the attribute maps based on the tile sizes, with larger tiles yielding less complex and clearer attribute maps.

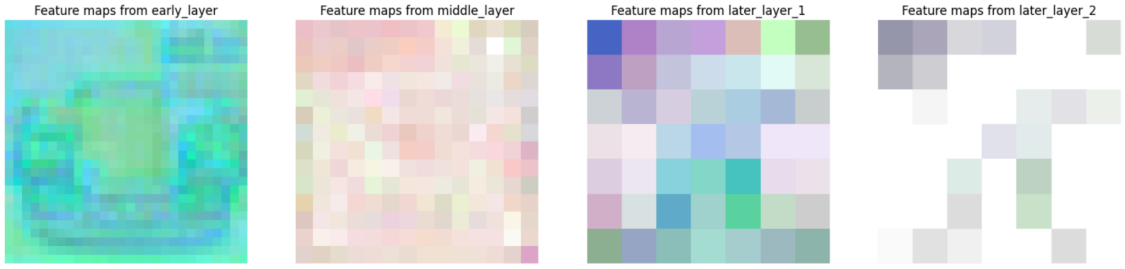
Pretrained Feature Maps - Baseline DenseNet



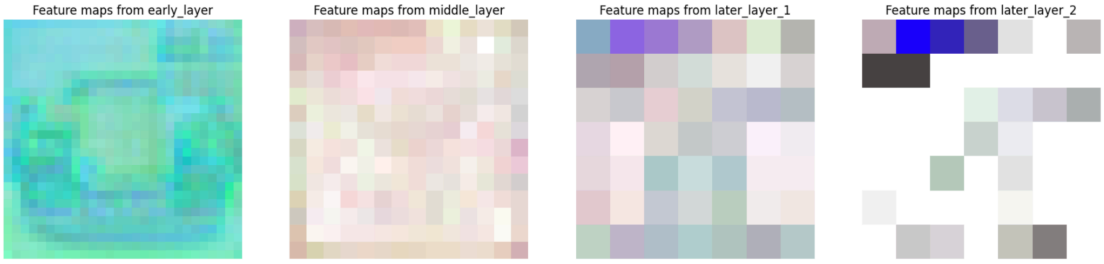
Trained Feature Maps - Baseline DenseNet



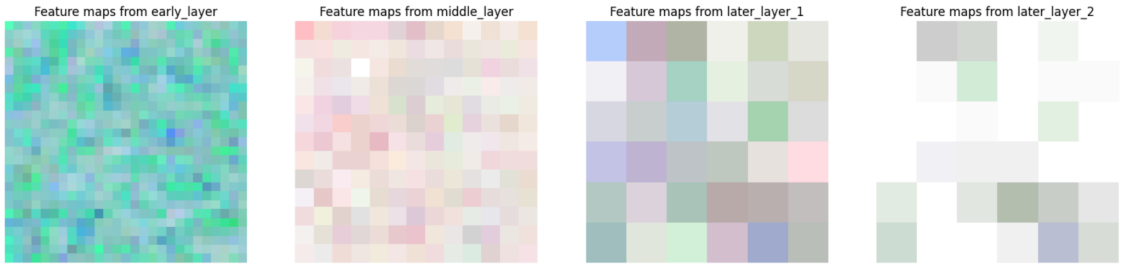
Pretrained Feature Maps - DenseNet 3x3



Trained Feature Maps - DenseNet 3x3



Pretrained Feature Maps - DenseNet 20x20



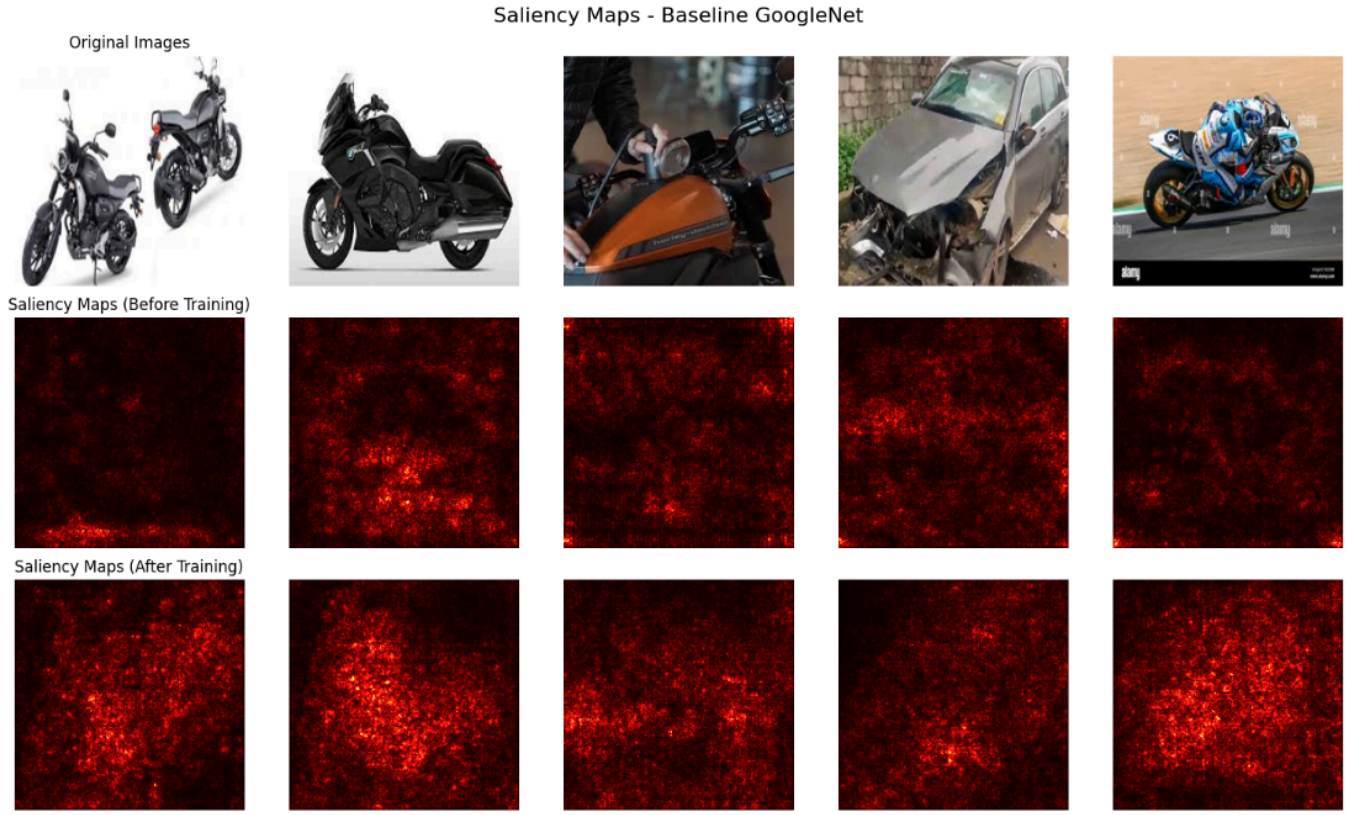
Trained Feature Maps - DenseNet 20x20

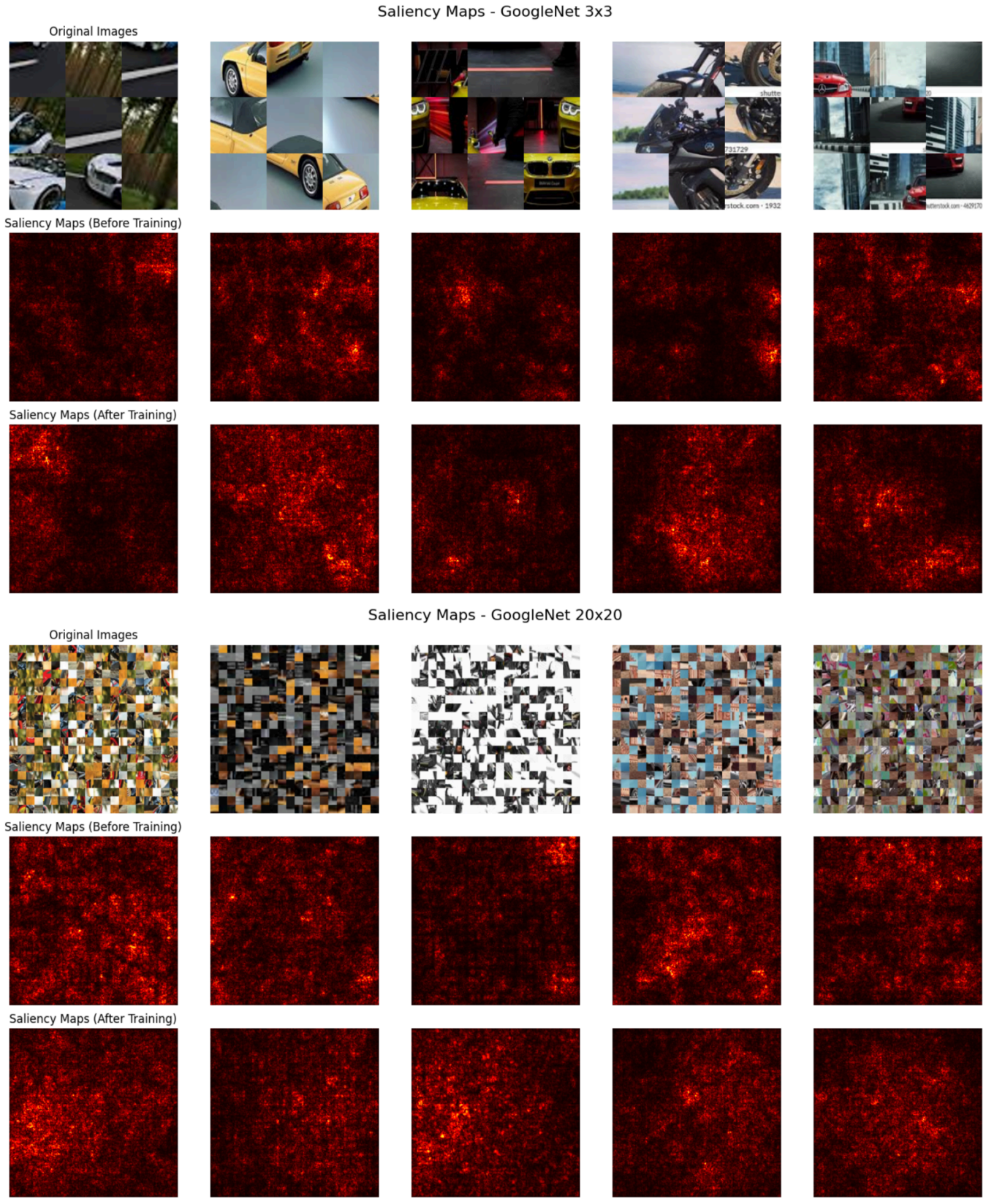
In summary, our investigation into feature maps across various tile resolutions revealed consistent layer behavior in both GoogleNet and DenseNet models, with notable distinctions primarily in deeper layers. Additionally, we observed that larger tile sizes led to less complex and clearer attribute maps, indicating the impact of resolution on interpretability.

4.3.2 Saliency Maps

The saliency maps of the GoogleNet model reveal notable distinctions between the pre-trained state and after specific data training, as well as across different tile sizes. Initially, before training, the saliency maps are uniformly activated, suggesting a generalized focus across the entire image space without distinguishing features of interest. Post-training, the saliency maps show more concentrated hotspots of activation, indicating a refined focus on relevant features within the images. This transition demonstrates the model's learning progression and increased specificity in identifying salient parts of the input.

As the size of the tiles changes, so does the model's attention distribution. For the baseline images, the saliency is relatively broad, which sharpens with training as expected with clear images. However, for the 3x3 and especially the 20x20 permutations, the pre-training maps appear diffuse, likely due to the model's initial unfamiliarity with the permuted structure of the inputs. After training, the saliency maps for these permutations become more defined, although not as pinpointed as in the baseline case, suggesting an improved but still challenging recognition of features within highly permuted tiles. This evolution underscores the model's ability to adapt its feature recognition capabilities to varying levels of input complexity through training.



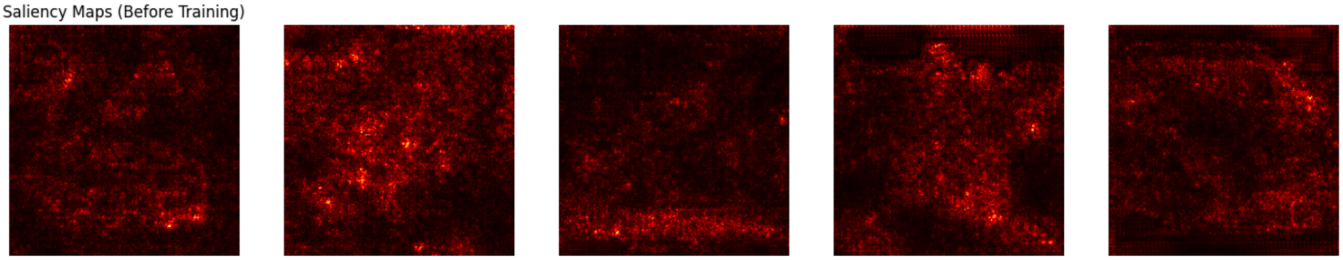


Similar to GoogLeNet, for DenseNet, the images appear to exhibit more spatial distortions before specific training, indicative of a broader interpretation of features. However, post-training, there is a noticeable refinement in detail accuracy. Additionally, the saliency maps exhibit more scattered patterns pre-training, while post-training, they tend to align more closely with the characteristics of each tile. Nevertheless, compared to the baseline, the saliency maps are noticeably less distinct.

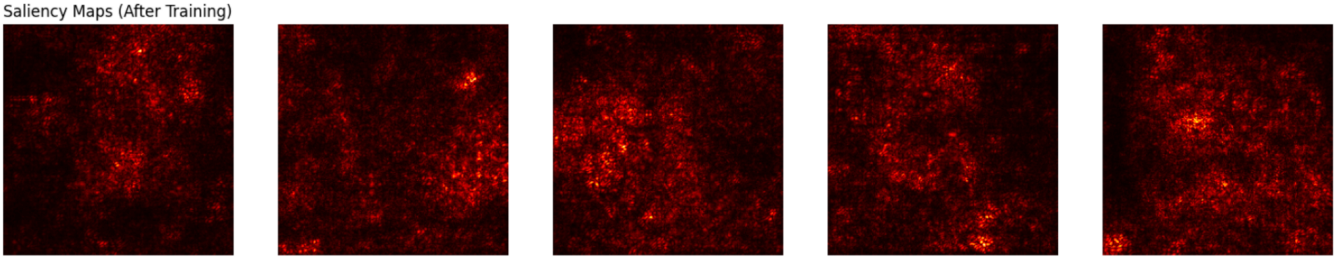
Saliency Maps - Baseline DenseNet

Original Images

Saliency Maps (Before Training)

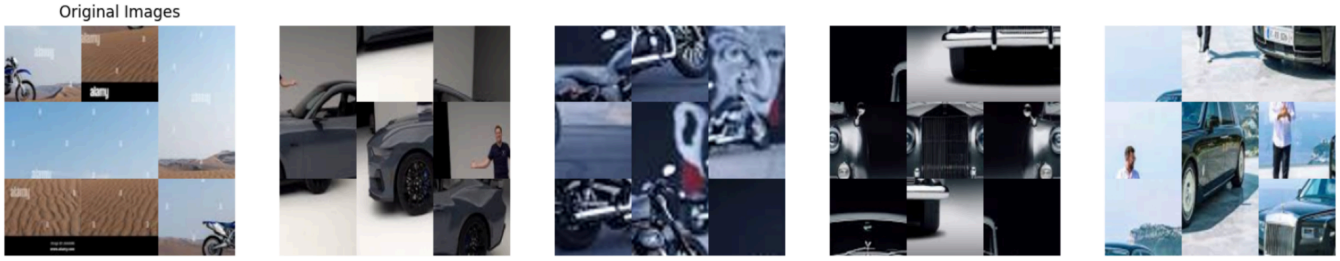


Saliency Maps (After Training)

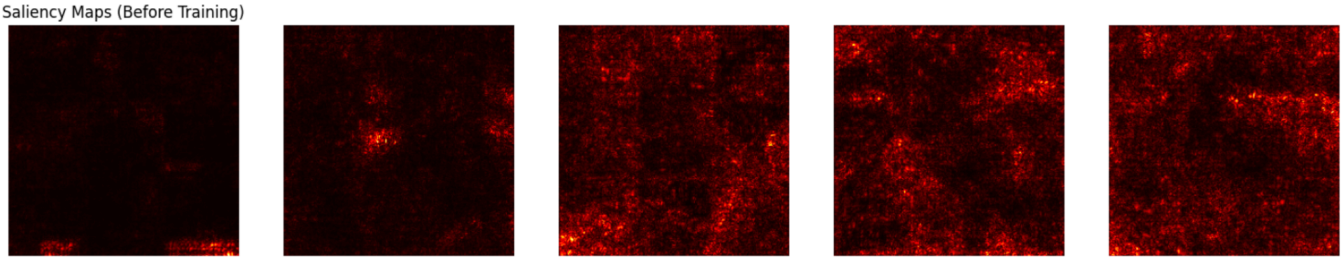


Saliency Maps - DenseNet 3x3

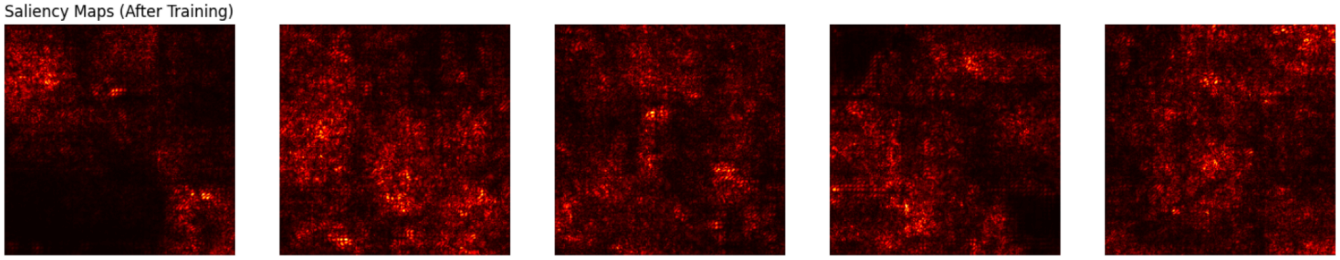
Original Images

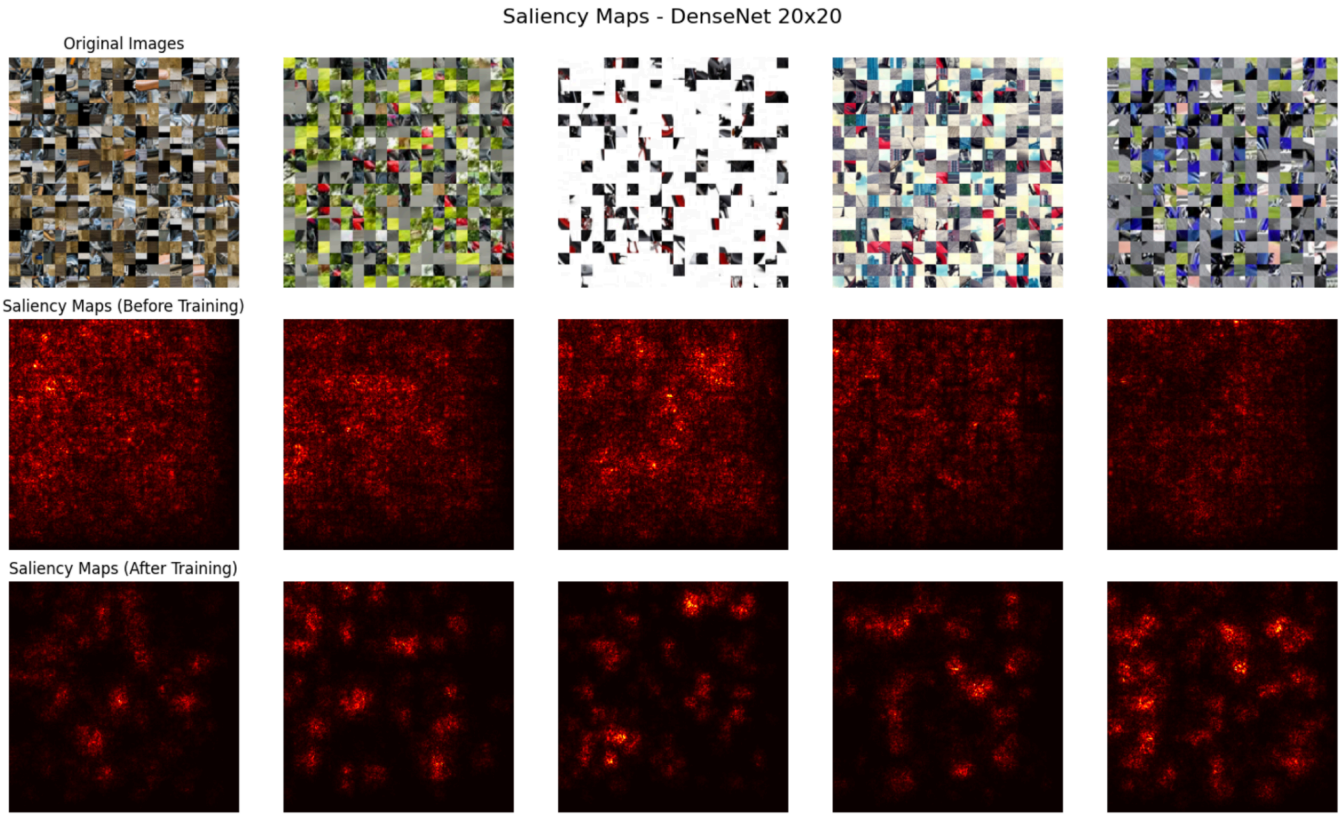


Saliency Maps (Before Training)



Saliency Maps (After Training)





In summary, both GoogLeNet and DenseNet demonstrate improved feature recognition and saliency map clarity post-training, indicating a refined focus on relevant image features. However, saliency maps for highly permuted tiles remain less distinct compared to baseline images even after training, suggesting ongoing challenges in identifying features within complex input structures.

4.4 Tile Resolution Impact - Summary

The exploration of tile resolution's impact on classification post-permutation reveals a fascinating interplay between model architecture and input complexity. As the granularity of permutations increases from baseline through to 20x20 tiles, both GoogleNet and DenseNet exhibit a remarkable adaptability, though with varying degrees of robustness. While pre-trained models begin with a generalized approach, post-training refinements hone their focus, sharpening saliency maps and feature activations to deal adeptly with permutation challenges. Notably, higher resolutions pose a significant strain, diluting feature clarity and complicating the classification task. GoogleNet's resilience is most pronounced, benefiting from lower learning rates and strategic augmentation, thus mitigating the disruption of intricate spatial relationships. Conversely, DenseNet showcases an optimal balance at moderate learning rates and reveals augmentation isn't a universal remedy, particularly when complexity overshadows inherent model strengths. These findings underscore the intricate balancing act required to maintain model accuracy amidst the disarray of permutations, highlighting the profound impact of tile resolution on the neural networks' ability to decipher and classify visual data.

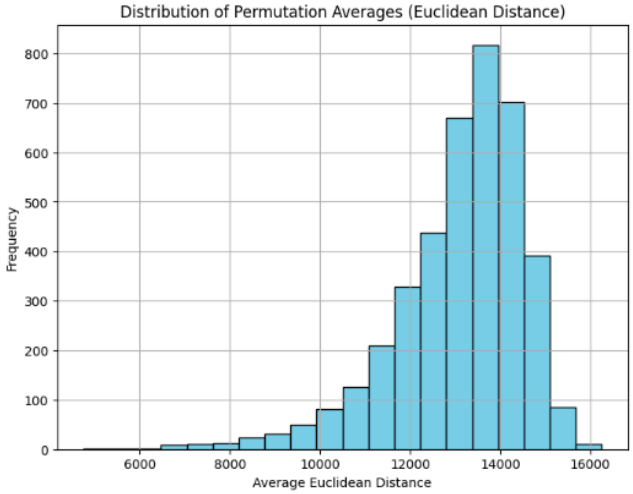
In summary, while we achieved notable enhancements across various permutations, there was a slight downward trend in model performance with increasing tile size, which became more pronounced at the 20x20 level.

5. Permutation Ranking-

In this section, we develop a method to rank permutations based on their impact on classification performance. We calculate the Euclidean distance for each tile from its original counterpart, chosen for

its ability to capture spatial relationships accurately. Summing these distances across all tiles within an image allows us to compute the average distance per image. We assume that greater distances indicate higher difficulty for classification by the model.

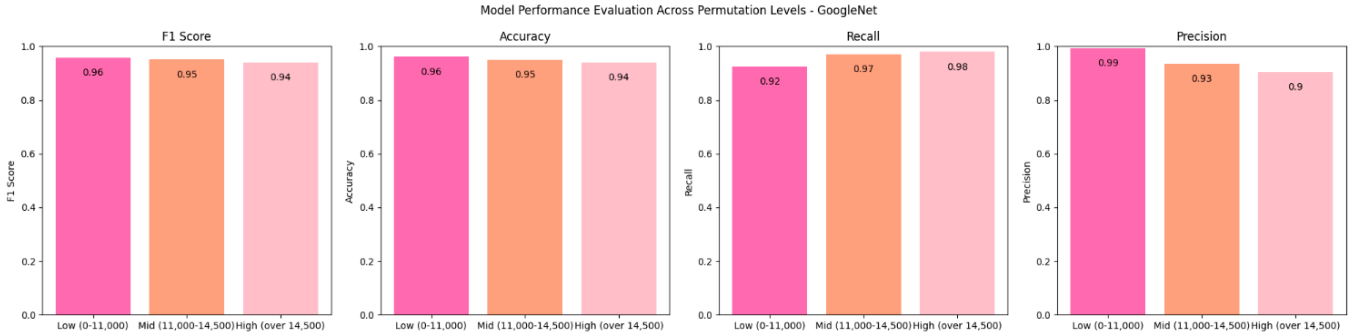
Initially, we analyze the distribution of average distances for random permutations:



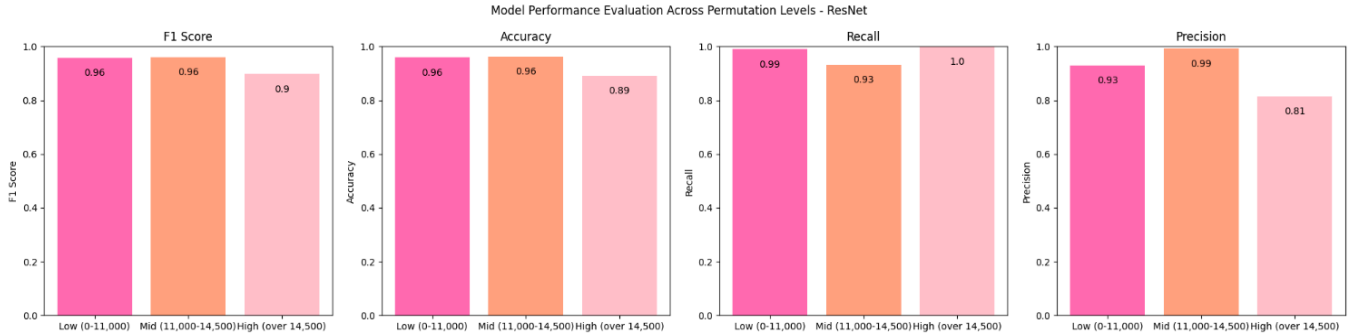
Based on this distribution, we categorize the transformations into three difficulty levels:

- Low: average distance below 11,000
- Mid: average distance between 11,000 and 14,500
- High: average distance above 14,500

Subsequently, we construct a dataset for each difficulty level, ensuring that each image falls within the corresponding average distance range. We then train the model and evaluate its performance on the validation dataset. The obtained results:



We observed a correlation between the average distance of permutations and the model's classification accuracy. As the average distance increases, indicating greater spatial distortion, we noticed a corresponding decrease in accuracy, albeit marginal. This suggests that more challenging permutations indeed pose difficulties for the classification model. Notably, the GoogLeNet architecture consistently demonstrated robust performance across various permutation types. However, it's worth noting that the disparity in performance may widen when applied to models with less favorable outcomes initially-ResNet.



Regarding the ResNet architecture, we observed relatively consistent performance across low and mid difficulty levels. However, there was a noticeable decrease in evaluation metrics for permutations categorized as high difficulty. This indicates that ResNet may struggle more with highly distorted images, highlighting the importance of considering model robustness when dealing with challenging permutations.

In summary, we developed a method to rank permutations based on their influence on classification performance, finding that greater spatial distortion led to decreased model accuracy. While GoogLeNet displayed consistent performance across permutations, ResNet struggled notably with highly distorted images. This underscores the importance of considering model robustness when dealing with challenging permutations.

6. References

[1] Frisk, M., Storgaard, K., & Gottfredsen, J. P. (2023). MULTI PRETEXT TASK SELF-SUPERVISED LEARNING FOR CULTURAL HERITAGE CLASSIFICATION. Aarhus University, Department of Computer Science.

7. Acknowledgments

To our ever-patient lecturer, who's been answering all our queries (and hopefully will award us a high grade).

To ChatGPT, for its unwavering emotional and professional support, and especially to the Promet "improve my english please"

To Google Colab Pro, for fueling our computational dreams (and not for every shekel spent on it).

And to everyone who endured the nail-biting wait until the finish line.

🙌