Deliverable 02

This report presents a comparative analysis of three post-hoc explanation techniques **Grad-CAM**, **AblationCAM**, and **ScoreCAM** used to visualize model attention in a ResNet50 model trained on ImageNet. We applied each method to ten diverse input images covering animals and object-centric scenarios, to explore how different CAM techniques localize discriminative image regions and how their internal mechanisms impact visual precision.

In well-framed animal images such as the **Goldfish** and **West Highland White Terrier**, all three methods were able to localize the target object. Grad-CAM consistently highlighted key semantic features like the *terrier's face* and the *goldfish's head*, though its activations tended to be broader, occasionally bleeding into surrounding regions. AblationCAM produced similar attention maps but often incorporated slightly more background, likely due to its reliance on channel-wise ablations. ScoreCAM, on the other hand, yielded the most precise and focused results across these cases, tightly isolating object contours with minimal noise and consistently excluding background distractors, such as humans or reflections.

Differences became more pronounced in complex or distant scenes. In the **Vulture** and **Flamingo** images, Grad-CAM and AblationCAM roughly localized the main bird but also included parts of the *chimney* or *water reflections*. ScoreCAM kind of outperformed both by generating object silhouettes with somewhat strong separation from non-target areas. This pattern repeated in the **American Coot** and **Kite** examples, where Grad-CAM and AblationCAM prioritized only the most visible bird with a bit of water reflection and the flowers, whereas ScoreCAM attempted to provide broader *but slightly noisier* contextual coverage. Such observations highlight ScoreCAM's strength in semantic expressiveness, though it may sometimes trade off selectivity for coverage in texture-rich scenes.

Object images like **Racecar** and **Orange** further reinforced this trend. All three methods identified the object accurately, but ScoreCAM demonstrated superior spatial coverage, highlighting fine-grained details such as the *car roof*, *hood*, and even *underbody shadow*, or cleanly isolating the *pulp region* of the orange while ignoring the bowl in the background. These findings suggest ScoreCAM's utility in precise object detection tasks, especially when the object occupies most of the frame.

In terms of computation, Grad-CAM is the most lightweight, requiring only a single backward and forward pass, it generates visualizations in approximately 2–5 seconds per image. AblationCAM, which performs 64–128 ablation passes by masking channels one-by-one, takes significantly longer (~1.5 to 3 minutes per image). ScoreCAM is the most computationally intensive due to its scoring of each activation map through separate forward passes, often taking multiple minutes. Therefore, while Grad-CAM is best suited for rapid analysis, ScoreCAM provides the highest-quality, context-rich visual explanations where runtime is not a constraint.

These findings confirm that while all three methods agree on object localization in principle, their implementations and computational strategies lead to varied performance across scene complexity, object proximity, and runtime constraints. Like how training data shapes feature learning (as seen in Task 1), the choice of visualization technique directly influences interpretability outputs. *Hence*, *selecting a CAM method must account for both the* ***nature of the image*** *and the* ***practical requirements of the interpretability task*****.**