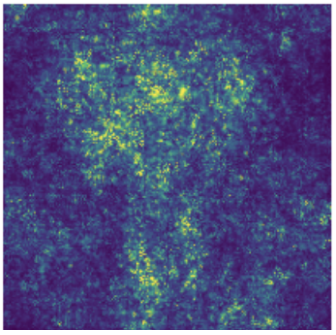
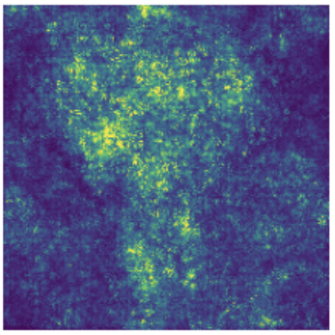
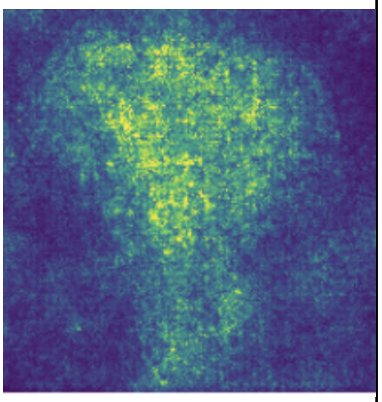
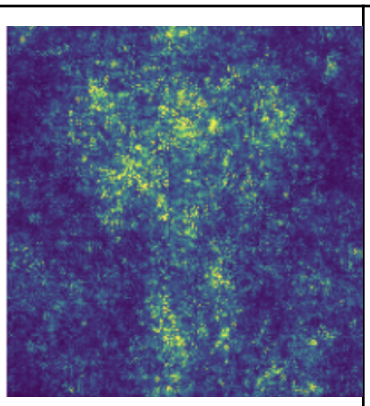
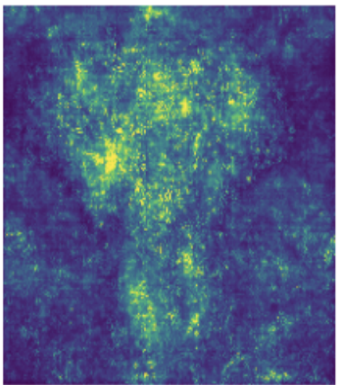
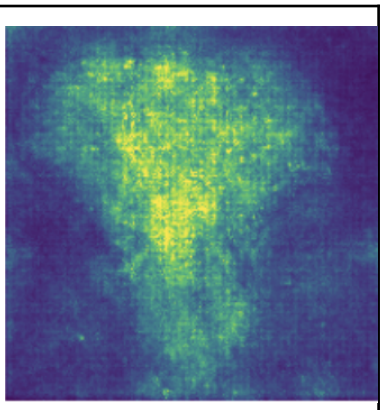
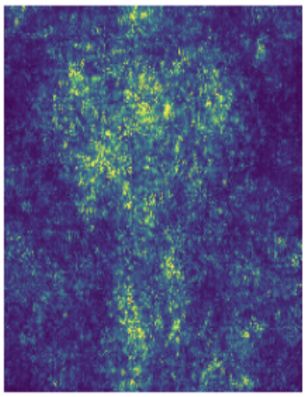
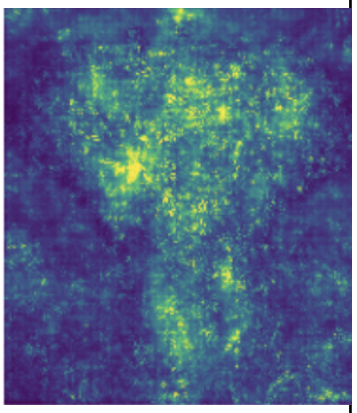
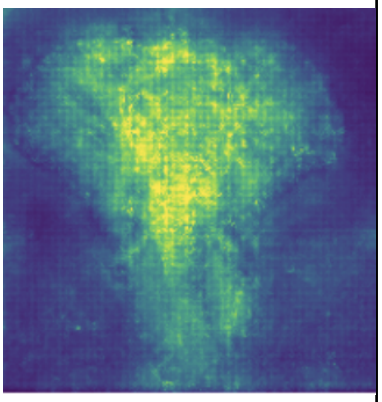
Abstract:

This report evaluates and compares multiple saliency visualization methods, Vanilla Gradients, Integrated Gradients, SmoothGrad, Occlusion, and Grad-CAM, to determine which are most worthwhile for the vision team moving forward. Among these, SmoothGrad and Integrated Gradients emerge as the most effective due to their clarity, stability, and alignment with human visual intuition. Both methods consistently highlight semantically meaningful features, such as the elephant's trunk or the dog's face, while reducing pixel-level noise and preserving fine details, making them superior for model interpretability and debugging. The report details experiments on varying SmoothGrad parameters (n_samples and stdev), occlusion configurations (window size and baseline), and analyses of multiple test images across five methods. It concludes with a counterfactual experiment that tests model sensitivity to image occlusion and summarizes the key qualities of a useful saliency map. This report assumes the audience possesses a working understanding of convolutional neural networks (CNNs), gradient-based attribution methods, and basic familiarity with visualization tools like PyTorch Captum, allowing technical yet accessible explanations of model interpretability techniques.

Smoothgrad parameters:

1.Choose one of the base images and create a 3x3 grid of saliency maps (top-1 assigned class only) by varying n_samples (5, 25, 125) and stdev(0.01, 0.05, 0.25).

	stdev = 0.01	stdev = 0.05	stdev = 0.25
n_samples = 5			

n_samples = 25			
n_samples = 125			

2.Describe how the saliency maps change (or don't).

As the number of samples (n_samples) increases from 5 to 125, the saliency maps become progressively smoother and less noisy, with clearer emphasis on the main features of the elephant's head and trunk.

When the noise standard deviation (stdev) increases, the maps become more diffused, small details fade, and the highlighted regions expand. At low stdev values (0.01, 0.05), the fine textures and edges are more visible, while at a high stdev (0.25), the attribution spreads widely, losing detail but emphasizing overall shape.

3.Choose one of the combinations to use for the rest of the report and provide justification for your choice.

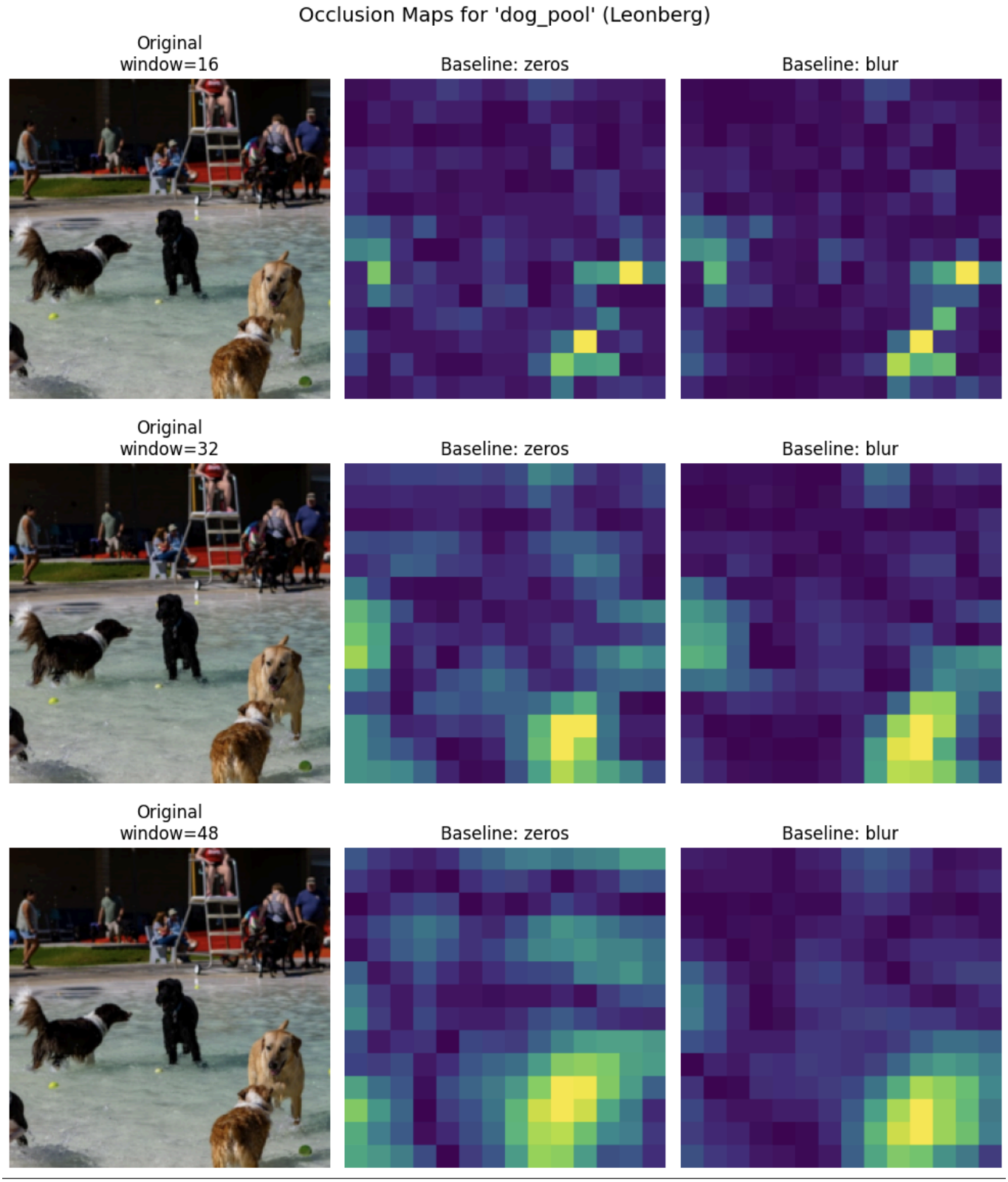
The best balance appears at n_samples = 25 and stdev = 0.05.

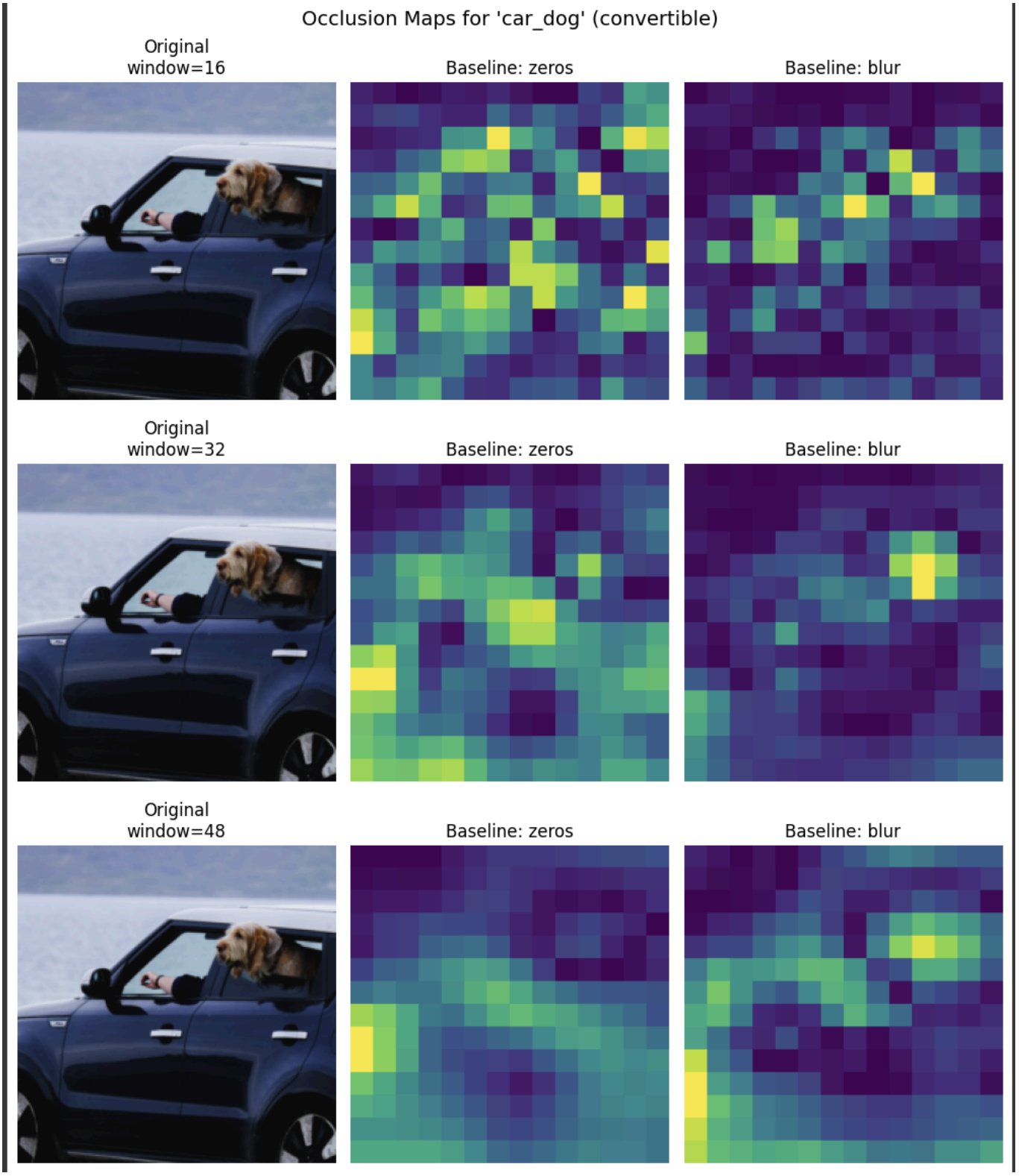
This combination produces a clear yet stable saliency map that captures the elephant's shape and main contours without excessive noise or over-smoothing. Lower sample counts (e.g., 5) show too much pixel-level noise, while very high stdev (0.25) overly blurs the structure.

Therefore, (n_samples=25, stdev=0.05) provides the most interpretable visualization for understanding where the model focuses when identifying the elephant.

Occlusion parameters

1.For each of the "dog_pool" and "car_dog" images, create a 3x2 grid of saliency maps (top-1 assigned class only) by varying window_size (16, 32, 48) and baselines (zeros and blurred) for the occlusion method. Leave the stride at 16.





2. Describe how the saliency maps change (or don’t)

For both “car_dog” and “dog_pool”, increasing the window size from 16 → 48 causes the occlusion maps to become smoother and less detailed.

- With smaller windows (16), the highlighted regions are more localized, showing fine-grained areas (like the dog’s head or car window) that strongly influence the prediction.

- As window size increases (32, 48), these activations become broader and blurrier, indicating that the model's sensitivity is being averaged over larger image regions.
- Comparing baselines, the blur baseline produces softer, more natural transitions, while the zeros baseline introduces sharper edges due to the stark contrast between occluded and visible areas.
- In both cases, the high-importance regions (yellow areas) correspond to the dog and nearby context, which are key to the model's classification.

3.Choose one of the combinations to use for the rest of the report and provide justification for your choice.

The optimal configuration is window_size = 32 with a blur baseline.
This setup offers the best trade-off between interpretability and stability, it highlights meaningful regions (like the dog's face or central object) clearly without excessive noise or overly broad attributions.
Smaller windows (16) produce fragmented, noisy patterns, while larger ones (48) lose detail and over-smooth.
The blurred baseline avoids artificial sharp edges seen in the zeros baseline and better represents how the model responds to natural feature occlusions.

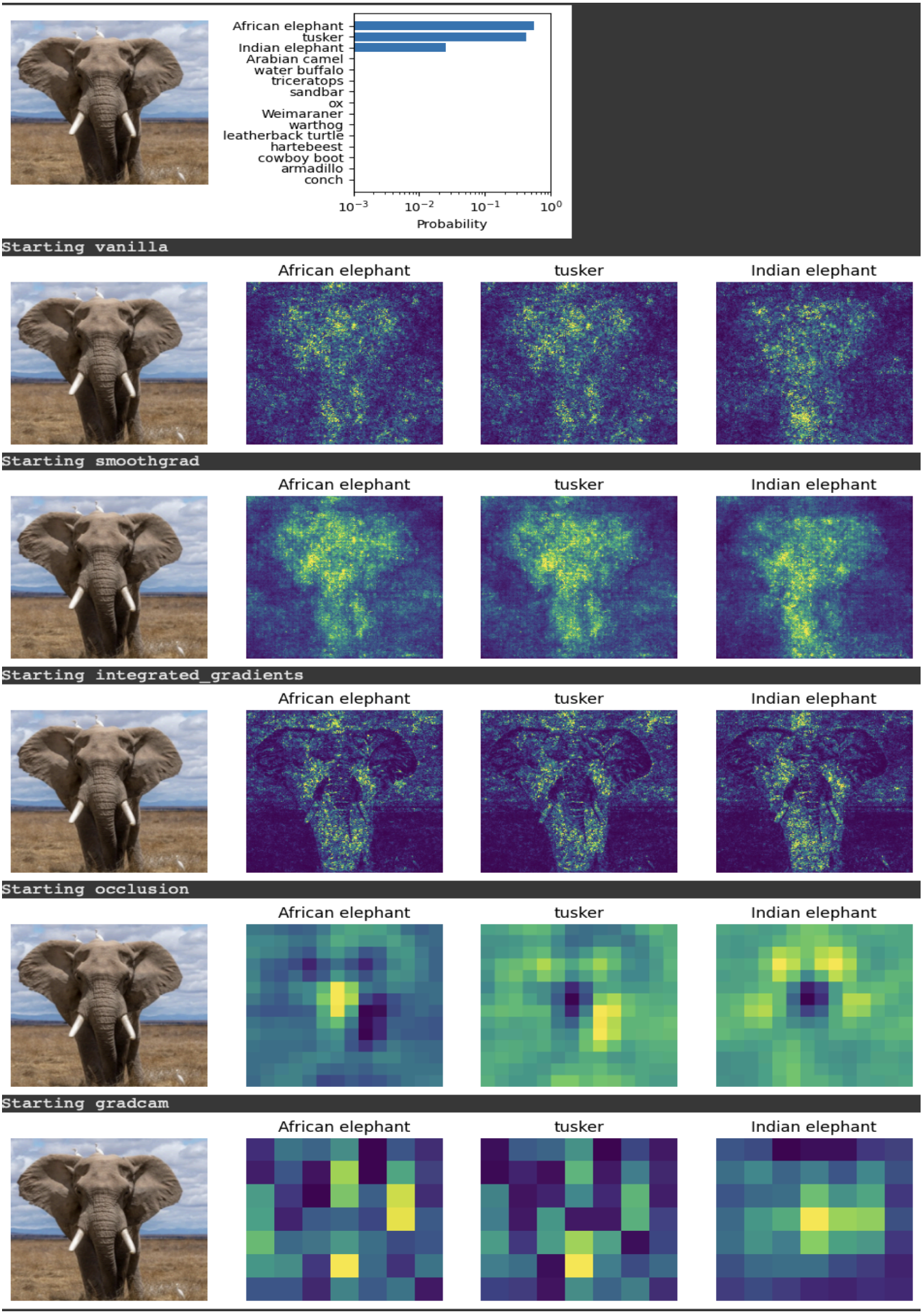
Explanations for the five images

1.Create saliency maps using the five methods for the top-3 predictions for each of the four base images. Include the top-15 probability bar chart.

2. In 1-2 paragraphs, note any saliency maps that appear sensible, and describe why.

3. In another paragraph, note any saliency maps that do not make sense, and describe why.

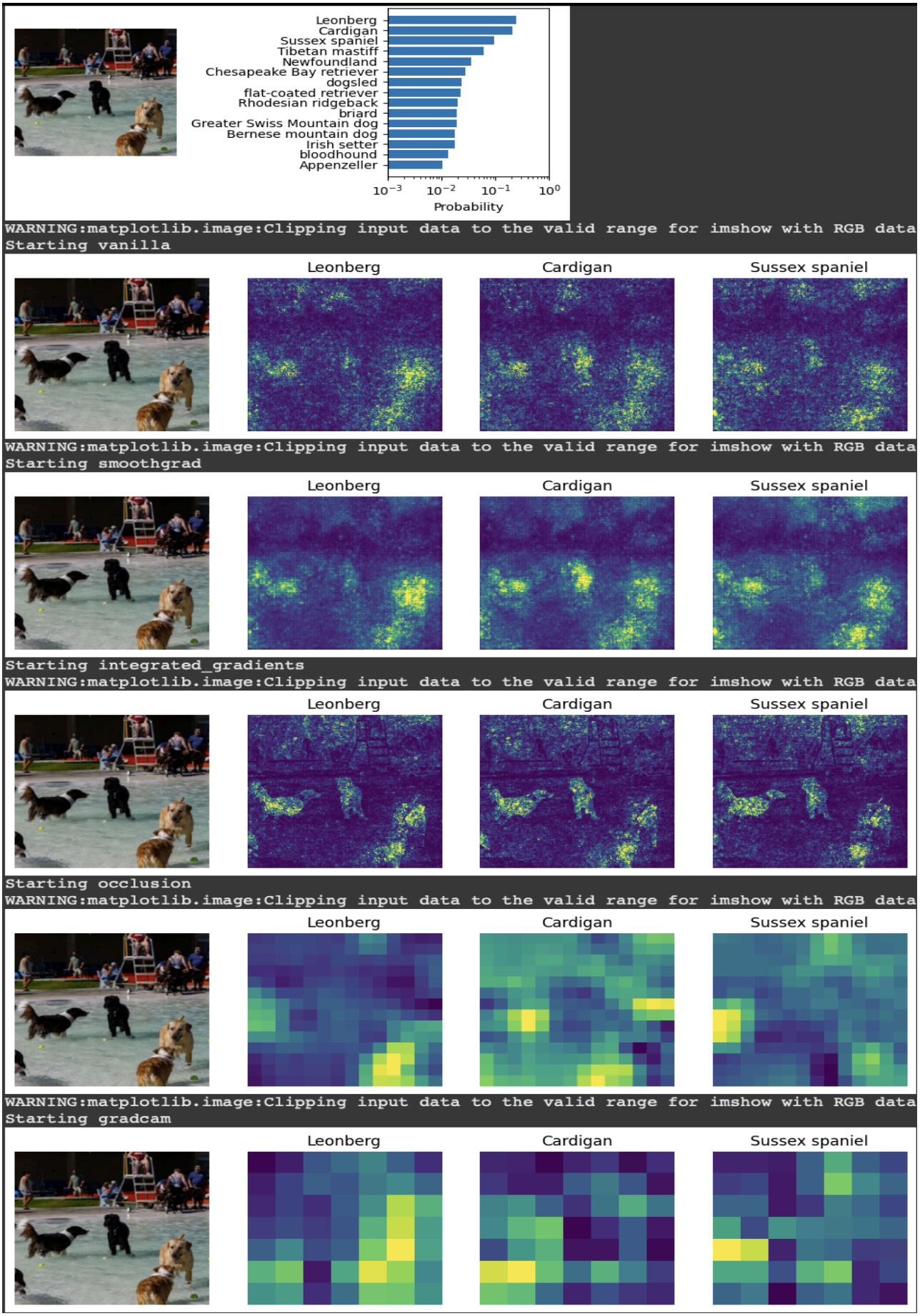
Image 1:


African elephant
tusker
Indian elephant
Arabian camel
water buffalo
triceratops
sandbar
ox
Weimaraner
warthog
leatherback turtle
hartebeest
cowboy boot
armadillo
conch
10^{-3}
10^{-2}
10^{-1}
10^{0}
Probability
Starting vanilla
African elephant
tusker
Indian elephant
Starting smoothgrad
African elephant
tusker
Indian elephant
Starting integrated_gradients
African elephant
tusker
Indian elephant
Starting occlusion
African elephant
tusker
Indian elephant
Starting gradcam
African elephant
tusker
Indian elephant

The saliency maps that appear most sensible are those produced by Integrated Gradients and SmoothGrad.
Both of these highlight key features of the elephant that are essential for classification, particularly the head, trunk, and ears. These regions correspond to distinctive features that differentiate elephants (especially between African and Indian species). The maps show consistent and dense activation around these areas, suggesting that the model's attention is directed toward semantically meaningful visual cues rather than background noise. The probability bar chart supports this, showing the highest confidence for "African elephant" and "tusker," both categories closely associated with the visible features emphasized in these maps.

On the other hand, the GradCAM and Occlusion maps appear less sensible. GradCAM produces blocky, coarse heatmaps with little clear focus on the elephant's defining features, implying poor spatial localization. The occlusion maps, though slightly more localized, still highlight inconsistent and sometimes irrelevant patches, failing to clearly capture the trunk or ears. The Vanilla gradients also appear noisy and scattered, lacking clear interpretability, indicating sensitivity to pixel-level variations rather than meaningful features. These less coherent maps are harder to interpret and do not align well with what a human observer would consider salient for identifying an elephant.

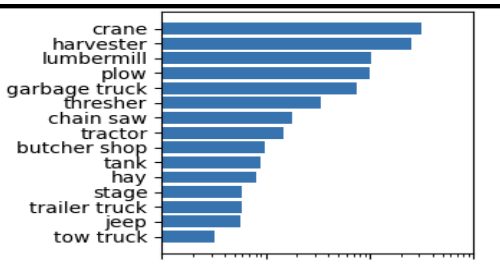
Image 2:


Leonberg
Cardigan
Sussex spaniel
Tibetan mastiff
Newfoundland
Chesapeake Bay retriever
dogsled
flat-coated retriever
Rhodesian ridgeback
briard
Greater Swiss Mountain dog
Bernese mountain dog
Irish setter
bloodhound
Appenzeller
10^{-3}
10^{-2}
10^{-1}
10^{0}
Probability
WARNING:matplotlib.image:Clipping input data to the valid range for imshow with RGB data
Starting vanilla
Leonberg
Cardigan
Sussex spaniel
WARNING:matplotlib.image:Clipping input data to the valid range for imshow with RGB data
Starting smoothgrad
Leonberg
Cardigan
Sussex spaniel
Starting integrated_gradients
WARNING:matplotlib.image:Clipping input data to the valid range for imshow with RGB data
Leonberg
Cardigan
Sussex spaniel
Starting occlusion
WARNING:matplotlib.image:Clipping input data to the valid range for imshow with RGB data
Leonberg
Cardigan
Sussex spaniel
WARNING:matplotlib.image:Clipping input data to the valid range for imshow with RGB data
Starting gradcam
Leonberg
Cardigan
Sussex spaniel

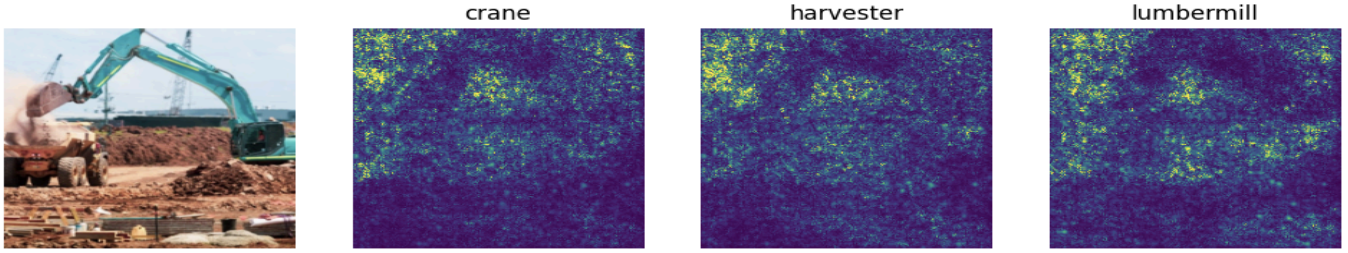
The saliency maps that appear most sensible are those generated by SmoothGrad and Integrated Gradients. Both clearly emphasize the bodies and heads of the dogs, which are the key distinguishing features for identifying breeds such as Leonberg, Cardigan, and Sussex spaniel. These regions align with where the model should focus, the fur texture, shape, and size of the dogs, rather than the background elements like the pool or people. The top-15 probability bar chart reinforces this interpretation: the model is most confident in dog-related categories (with “Leonberg” having the highest probability), and the highlighted regions in these maps visually justify that confidence.

In contrast, the Vanilla Gradient, Occlusion, and GradCAM maps appear less meaningful. Vanilla gradients are noisy and scattered, showing no clear focus on the dogs’ main features. The occlusion maps have blocky patches that often highlight irrelevant areas, including parts of the background, which do not contribute to breed classification. Similarly, GradCAM’s low-resolution grid-like activations lack detail and do not correspond to the dogs’ key features. These limitations make the latter maps harder to interpret and less reliable for understanding how the model distinguishes between similar dog breeds.

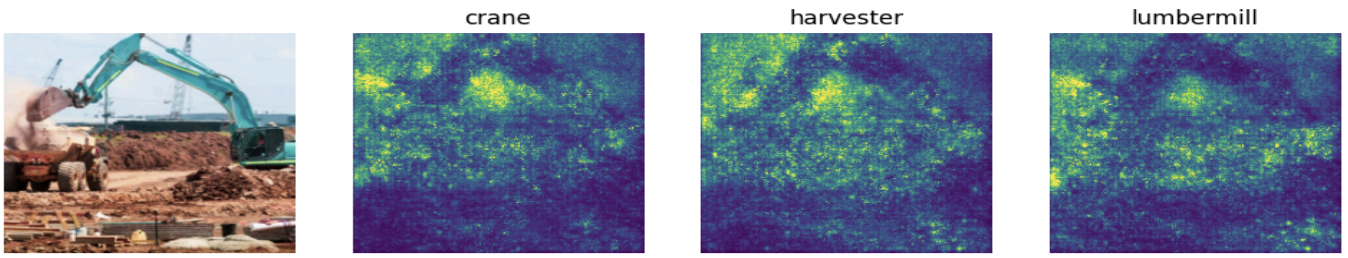
Image 3:



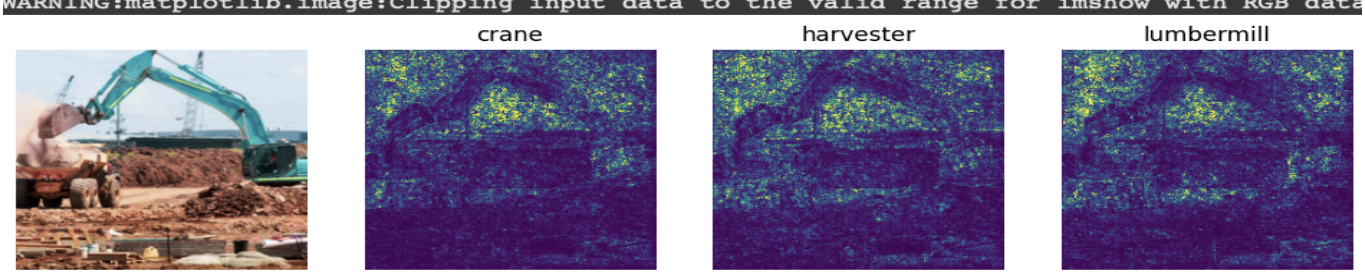

```
WARNING:matplotlib.image:Clipping input data to the valid range for imshow with RGB data
Starting vanilla
```



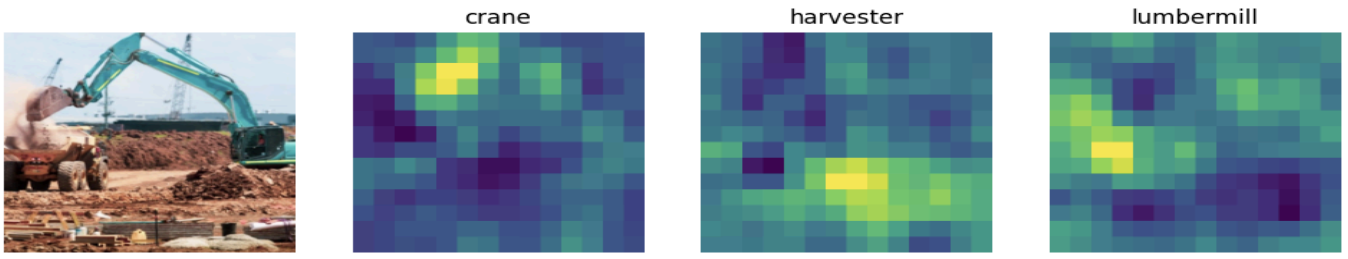

```
WARNING:matplotlib.image:Clipping input data to the valid range for imshow with RGB data
Starting smoothgrad
```



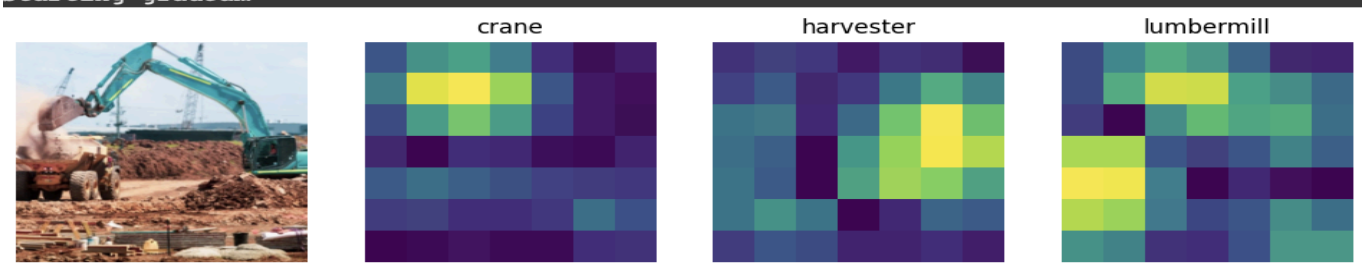

```
Starting integrated_gradients
WARNING:matplotlib.image:Clipping input data to the valid range for imshow with RGB data
```




```
Starting occlusion
WARNING:matplotlib.image:Clipping input data to the valid range for imshow with RGB data
```




```
WARNING:matplotlib.image:Clipping input data to the valid range for imshow with RGB data
Starting gradcam
```

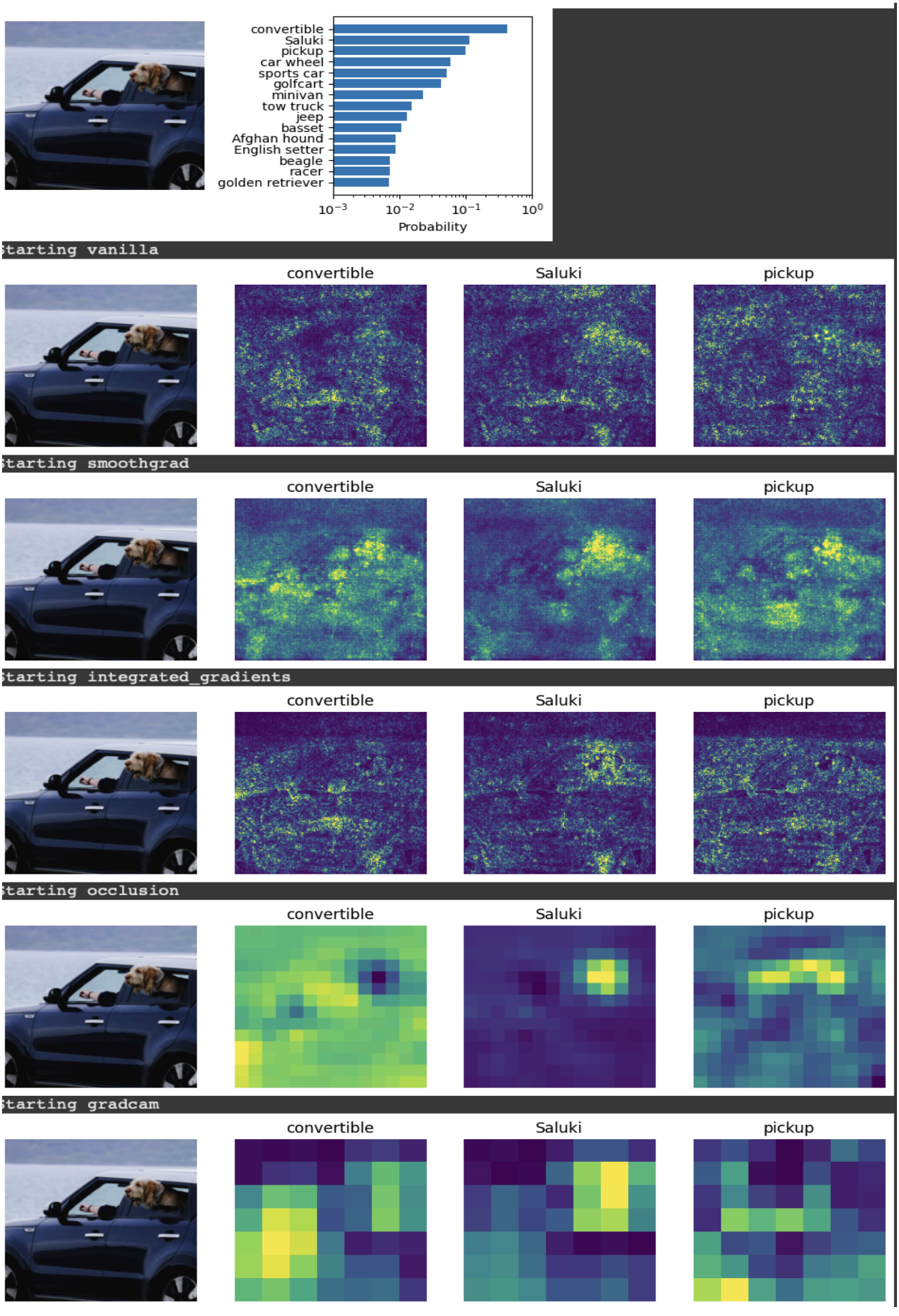


The top-15 probability bar chart shows that the model is most confident the object is a crane, followed by harvesterand lumbermill, with smaller probabilities for classes like *garbage truck*, *chainsaw*, and *tow truck*. This ranking makes sense given the image, a large industrial machine on a construction-like site with dirt and logs which visually aligns best with heavy-equipment categories such as cranes or harvesters.

Among the saliency maps that appear sensible, the SmoothGrad and Integrated Gradients methods stand out. In both, the highlighted (bright) regions correspond to the arm and bucket of the excavator and parts of the machinery body, which are visually distinctive and relevant to predicting “crane.” These methods show consistent focus on the functional parts of the machine, suggesting the model is attending to meaningful visual cues rather than background noise. The Grad-CAM results also look coherent, showing concentrated activation around the machinery rather than the dirt background, indicating spatially focused reasoning.

The less sensible maps are from vanilla gradients and occlusion. The vanilla saliency outputs are noisy, with bright spots scattered throughout the image rather than on the machinery, implying the raw gradients are unstable and less interpretable. The occlusion maps are overly coarse, producing blocky attention patterns that miss key structural details , for instance, they emphasize random patches of dirt or sky instead of the crane’s moving parts. This limits their interpretability and suggests these methods do not accurately reflect the model’s true decision process for this image.

Image 4:


convertible
Saluki
pickup
car wheel
sports car
golfcart
minivan
tow truck
jeep
basset
Afghan hound
English setter
beagle
racer
golden retriever
10^{-3}
10^{-2}
10^{-1}
10^{0}
Probability
Starting vanilla
convertible
Saluki
pickup
Starting smoothgrad
convertible
Saluki
pickup
Starting integrated_gradients
convertible
Saluki
pickup
Starting occlusion
convertible
Saluki
pickup
Starting gradcam
convertible
Saluki
pickup

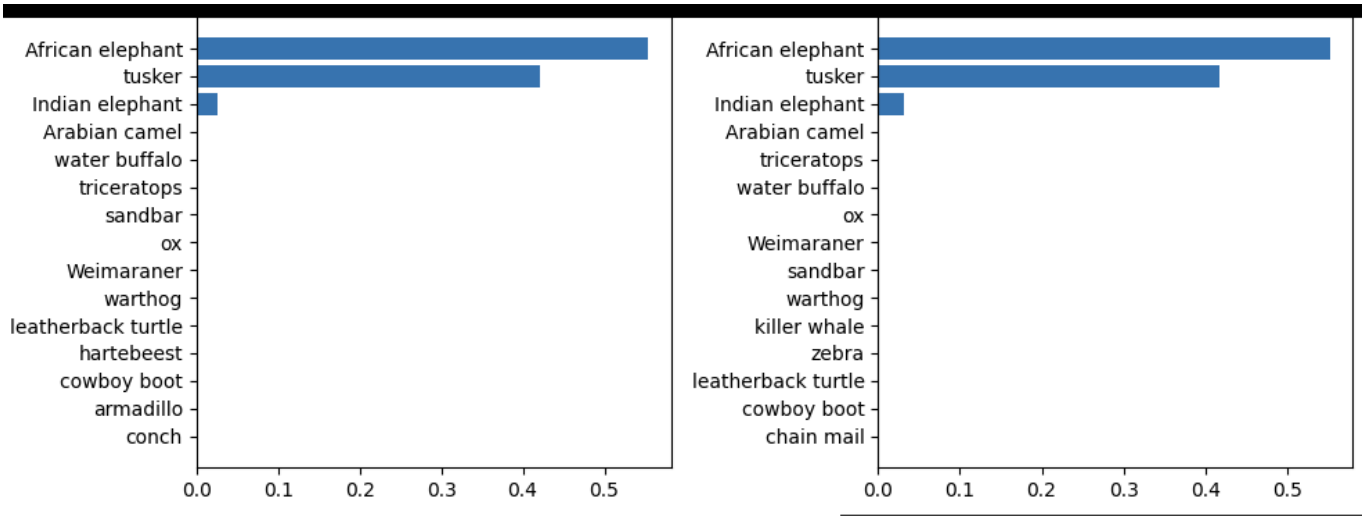
The top-15 probability bar chart shows that the model assigns the highest probabilities to “convertible”, “Saluki”, and “pickup”, followed by related classes such as *car wheel, sports car, minivan,* and *tow truck*. This indicates that the model is uncertain whether the main subject is a car type or a dog breed, which makes sense given that the image depicts a dog (possibly a Saluki) sitting in a convertible-type car.

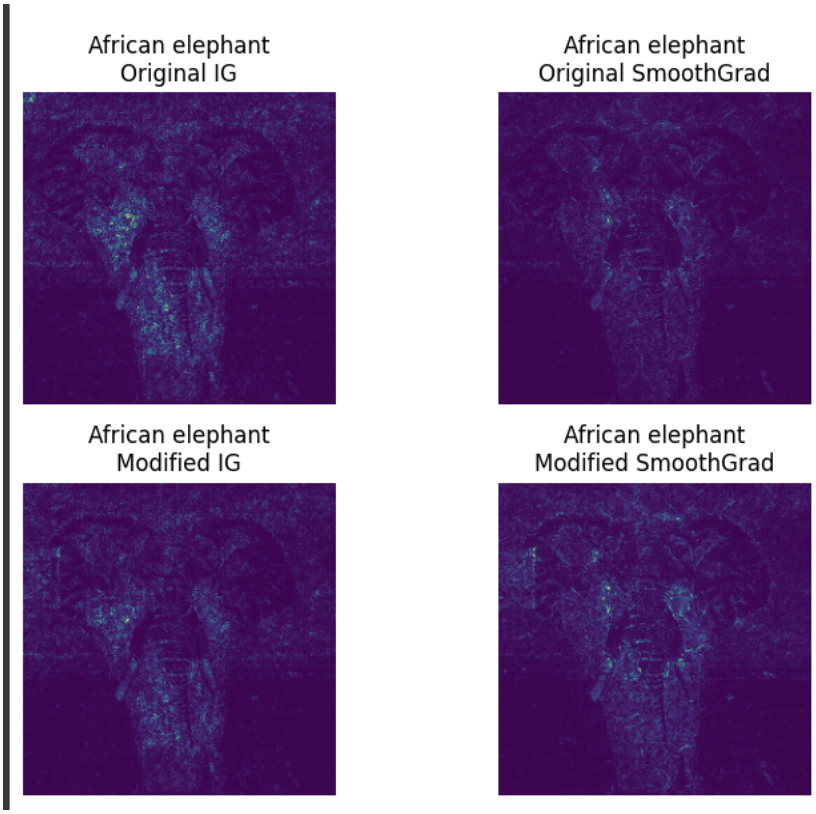
Among the saliency maps, the SmoothGrad and Integrated Gradients maps appear the most sensible. For both “convertible” and “Saluki”, these maps highlight meaningful regions, specifically, the front of the car and the dog’s head, which are the key distinguishing features for those classes. SmoothGrad effectively reduces noise and emphasizes coherent visual areas, while Integrated Gradients balances between the car structure and the animal, suggesting the model considers both objects in its decision.

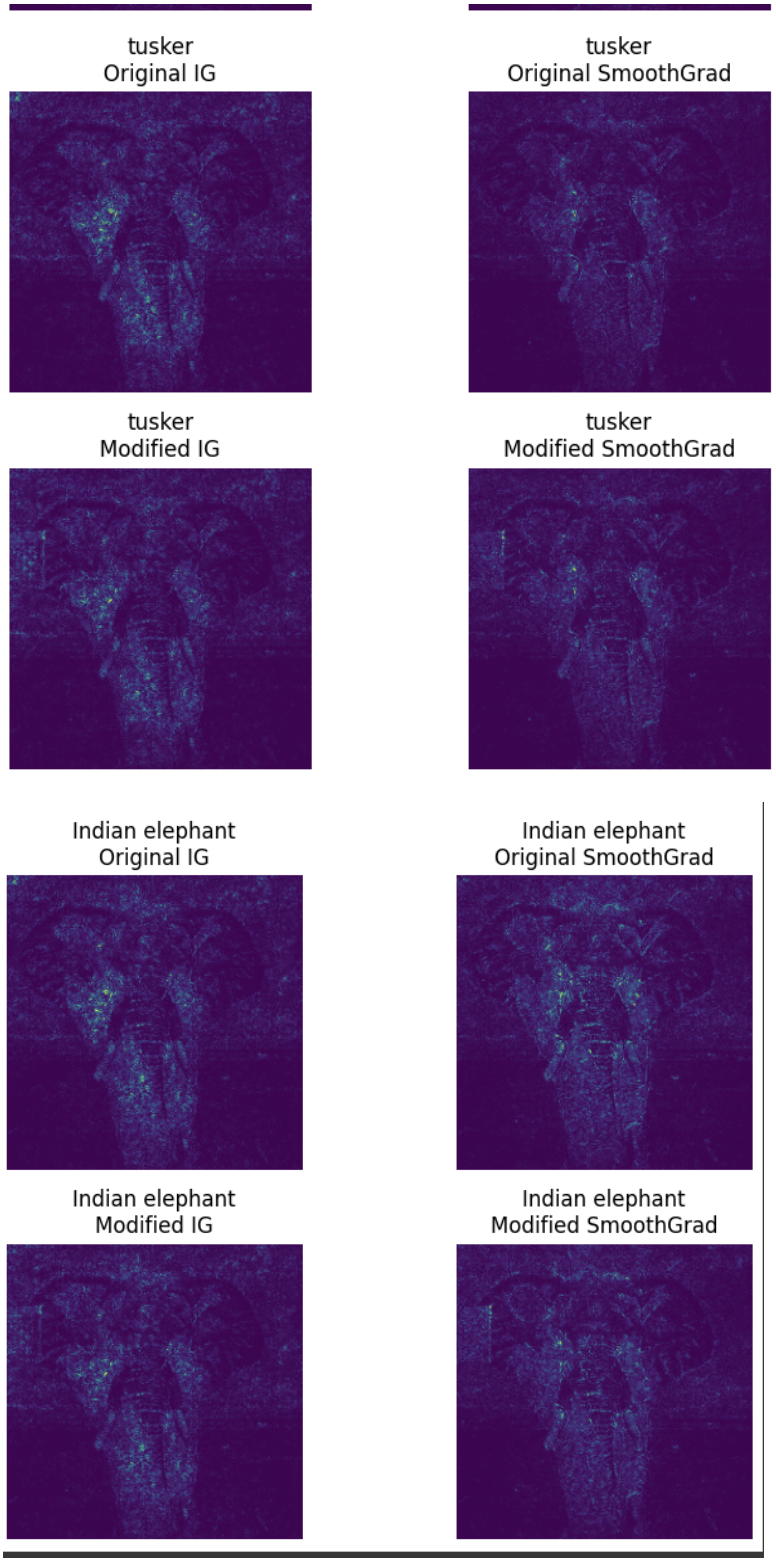
On the other hand, the Vanilla, Occlusion, and Grad-CAM maps are less interpretable. Vanilla gradients appear noisy and diffuse, offering little insight into specific features influencing classification. Occlusion maps seem overly coarse, focusing on broad patches that don’t clearly correspond to semantic regions of the image. Similarly, the Grad-CAM outputs appear too low-resolution and inconsistent across classes, missing finer details like the dog’s head or car edges. Hence, while some maps sensibly capture relevant cues (SmoothGrad and Integrated Gradients), others fail to produce clear or interpretable attributions for the model’s predictions.

Create an image counterfactual

1.Modify one image (e.g., occlude a region, paste a simple shape, recolor a part). Show before/after Top-15 probabilities and saliency maps for 1-2 methods of your choice, for the top 3 class assignments.







2.Explain whether the changes match your expectations.

After occluding part of the elephant image, the Top-15 probability plots show that the model still correctly classifies the image as an African elephant, but its confidence in the top classes changes slightly.

- The probability for *African elephant* and *tusker* dropped slightly compared to the original, showing that the occluded region (near the elephant's head and trunk) contains features that the model relies on for its decision.

- The rankings of the next few classes (like *Indian elephant* and *Arabian camel*) remain similar, meaning the model's overall understanding of the image context is consistent, but with reduced certainty.

  The saliency maps (both Integrated Gradients and SmoothGrad) visually confirm this change.

- Before occlusion, both methods strongly highlight the trunk, ears, and tusks distinctive features used to identify an elephant.
- After occlusion, the saliency intensity weakens around the blacked-out region and slightly shifts to the remaining visible parts (ears and body).
- SmoothGrad produces smoother, more stable patterns, while Integrated Gradients shows sharper activation boundaries.
  These patterns indicate that the occluded region contained discriminative features, and the model appropriately redirected its focus when that information was removed.

Takeaway

Based on your analyses, what makes a saliency map useful or interpretable? What would you want from an ideal explanation method? (1 paragraph)

A saliency map is most useful and interpretable when it clearly highlights the image regions that directly influence the model's decision in a way that aligns with human intuition. It should emphasize meaningful, task-relevant features, such as the identifiable parts of an object, rather than scattered or noisy regions. An ideal explanation method would combine clarity, consistency, and faithfulness to the model's reasoning, producing visualizations that are stable across small input variations and easy for humans to interpret without requiring deep technical knowledge. Ultimately, it should provide a transparent bridge between the model's internal processing and human understanding of *why* a certain prediction was made.

AI usage documentation

AI assistance (ChatGPT, GPT-5) was used to help me understand concepts related to saliency maps, visualization techniques, and explanation methods. It supported my learning process by clarifying definitions and guiding how to structure my written responses. However, all analysis, interpretation, and final explanations of the saliency results were done independently based on my own understanding and observations.

Code

https://colab.research.google.com/drive/1j6XdeaxEVkraIpYrT6xN6PLXOxvGJasi?usp=sharing