SWAG: Superpixels Weighted by Average Gradients for Explanations of CNNs
Supplementary Material

Thomas Hartley
Cardiff University
hartleytw@cardiff.ac.uk

Kirill Sidorov
Cardiff University
sidorovk@cardiff.ac.uk

Christopher Willis
BAE Systems Applied Intelligence
chris.willis@baesystems.com

David Marshall
Cardiff University
marshallad@cardiff.ac.uk

November 10, 2020

1 Introduction

This supplementary material contains elements that were withheld from the main paper due to space constraints. The first is a brief discussion on the hyper-parameter search used to determine the $\text{SWAG}_{\text{I+G}}$ hyper-parameters, this can be found in Section 4. Secondly we provide many more qualitative examples for explanations generated with both VGG16 and ResNet50 for ImageNet, Stanford Dogs, CUB Birds, and Oxford Flowers. These can be found in Sections 6.1, 6.2, 6.3, and 6.4 respectively. As with the main paper we recommend viewing these images in colour.

2 SWAG Pooling Method

In this section we quantitatively justify the method chosen for pooling the gradients into superpixels. In the main paper we proposed taking the mean of the gradients within a superpixel as the importance score for that region. However there are multiple other methods that could have been chosen. We therefore run the local accuracy experiment with a number of substitutions for scoring using the mean value. These alternatives are: minimum, maximum, ℓ1-norm, ℓ2-norm, variance, standard deviation, and the sum. The results from the local accuracy experiment can be found in Table 1. Here we confirm our initial intuition that taking the mean of the gradients within a superpixel gives the best local accuracy performance for both networks.

3 Alternative View of Local Accuracy Charts

Additional views of the local accuracy charts can be seen in Figure 1. These allow a clearer view of how the different methods for generating explanations compare to each other.

Method	VGG16	ResNet50
ℓ1-norm	0.099	0.124
ℓ2-norm	0.096	0.122
Maximum	0.100	0.125
Mean	**0.092**	**0.119**
Minimum	0.095	**0.119**
Standard Deviation	0.097	0.124
Sum	0.099	0.124
Variance	0.097	0.124

Table 1: AUC results for the local accuracy experiment. Lower is better. Here we see that of the pooling methods tried, taking the mean works the best.

4 SWAG$_{\mathbf{I+G}}$ Hyper-parameter Grid Search

With the introduction of SWAG$_{\mathrm{I+G}}$ there are now 3 weights, w_s, w_c, and w_g that allow control over the spatial, colour and gradient components respectively. The values assigned to these control the amount of influence each component exerts on the formation of the superpixels. In the original method, w_s is fixed in such a way as to try keep each superpixel with a set distance of a superpixel centre. Here $w_s = \sqrt{N/K}$, where N is the number of pixels in the image, and K is the desired number of superpixels. We do not change this for SWAG$_{\mathrm{I+G}}$. The authors of the SLIC found that setting w_c to 10 gave them the best results for segmenting an image into superpixels. This involved tests such as seeing how well the superpixel boundary aligns with a human annotated image. As we have scaled our gradient to produce values within a similar range as the colour components, we initially set w_g to 10 as well.

We perform a grid search over the w_g and w_c values from 4 to 20 in steps of 4. Assuming a centre point of 10, approaching a value of 20 will diminish that element, whilst approaching 4 will increase the prominence in the superpixel generation method. We chose to cap the search at a maximum value of 20 so as to still maintain a link between the superpixels and the underlying image. An example of the effect on superpixels that varying these values has is shown in Figure 2. To evaluate how a chosen set of hyper-parameters is performing we run the local accuracy experiment from RISE. We perform our grid search on the Stanford Dogs dataset using a trained ResNet50 model. We found that a w_c value of 20 and a w_g value of 8 gave best results, suggesting that a superpixel better explains a region when it can escape the constraints of the RGB image and extend over the natural image boundaries. In Figure 3 we show the interplay between the two values when the other is fixed. It is interesting to note that the local accuracy continues to improve as the influence of the colour space component is diminished, however we felt that reducing it too much would ‘detach’ the explanation from the underlying image. However, it would be interesting to explore this further in future work.

5 Weak-Localisation Details

We produce a bounding box by thresholding the explanation heatmap (sweeping through a range of thresholds). As per [2] we report results using three simple thresholding methods. First we threshold the pixel value (scaled to between 0 and 1) in steps $[0 : 0.05 : 0.95]$. Second, we threshold using the mean intensity of the heatmap (μ_I) by taking our threshold $\alpha \in [0 : 0.5 : 10]$ and multiplying by the mean to form $\alpha\mu_I$. Finally, we threshold using the energy of the heatmap in such a way that the most salient region of the heatmap covers a defined percentage of the energy, where $\alpha \in [0 : 5 : 95]$.

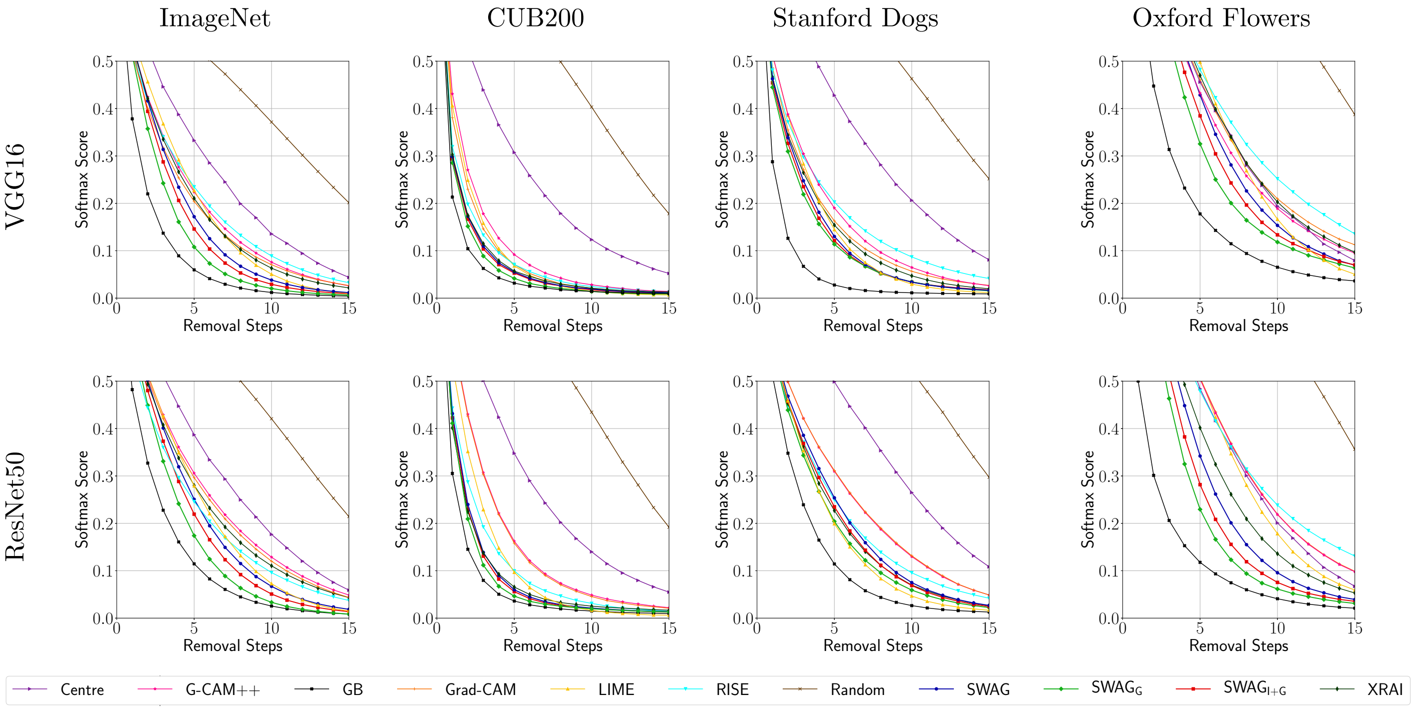

Figure 1: Zoomed in local accuracy AUC charts. Best viewed in a PDF viewer with zoom ability.

6 Additional Qualitative Results - Images

Over the following pages we present additional examples of our proposed methods compared to Grad-CAM [5], Grad-CAM++ [1], LIME [4], and RISE [3]. We separate each dataset into their own section. Every image has been classified correctly by the networks.

6.1 ImageNet

In Figure 4 and Figure 5 we show examples taken from the ImageNet validation set. There are a number of features that make SWAG, SWAG$_{\mathrm{I+G}}$, and SWAG$_{\mathrm{G}}$ desirable compared to the other methods tested. In particular we note how fine our explanations are compared to much coarser CAM based methods, This is presrumably one of the core reasons we achieve better local accuracy. For example, in the Black Grouse class in Figure 4, note how much of the area surrounding the object is highlighted in the CAM methods, and RISE. Compared to these we see the precision with which our method can highlight the areas really important to the network. We also note the similarities between RISE and our method. For example, both methods highlight small regions of the image in the Chainsaw and Washbasin classes in Figure 5 despite only requiring a single pass through the network compared to RISE's 4,000 − 8,000. A possible downside to our technique is that in some cases it can produce an explanation that is noisy. This can be seen in the Barbershop class, in particular for the VGG16 network. Although our explanation seems to accurately locate the barber pole and sign, there are a number of additional superpixels that are highlighted which could lead to confusion.

6 ADDITIONAL QUALITATIVE RESULTS - IMAGES

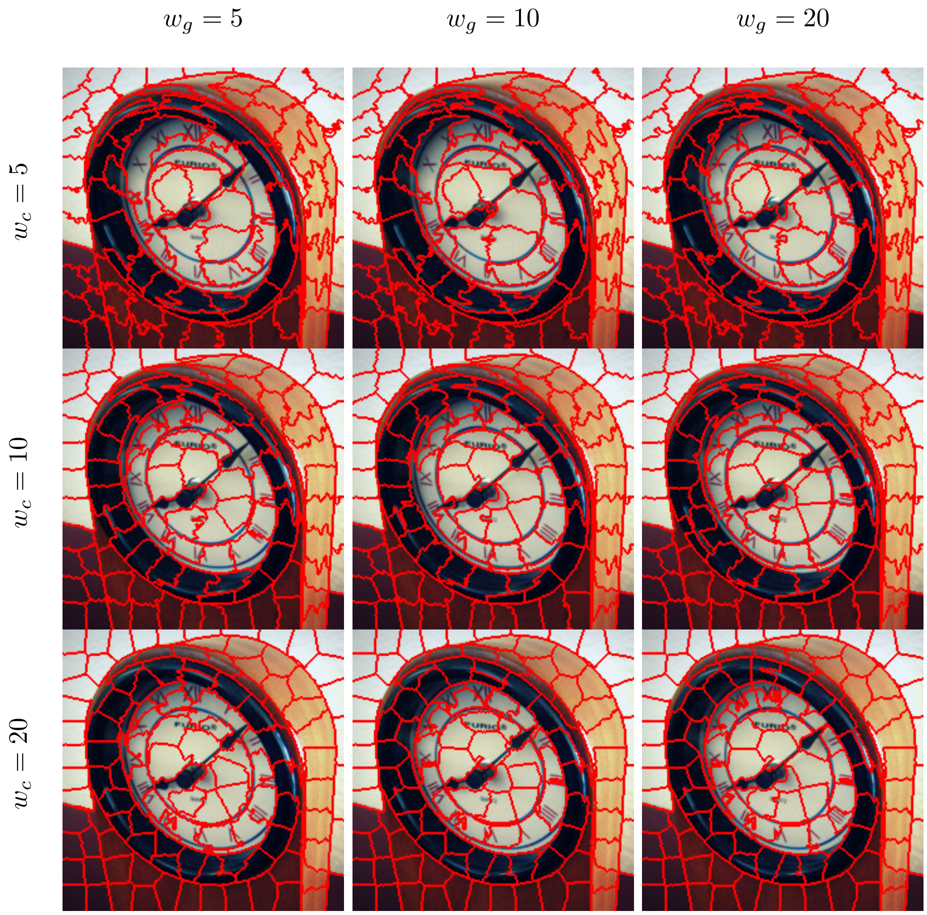

Figure 2: Comparison showing how altering the balance of w_g and w_c effects the generation of superpixels. For ease of viewing we have reduced the number of superpixels generated to 150

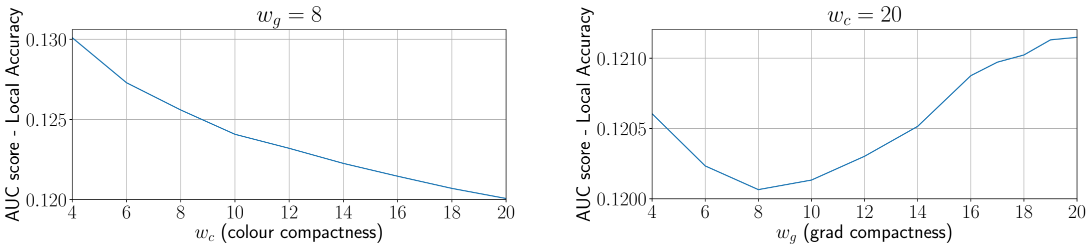

Figure 3: Grid Search results showing the interplay between each parameter when we fix each at their best values.

6.2 Stanford Dogs

Figure 6 contains five examples from the Stanford Dogs dataset. In this fine-grained dataset we see that our method is able to accurately identify small regions on the dogs body that the networks find important. In particular this seems to be the face region of the dog. This is corroborated by the other explanations, however, note the coarseness of the CAM methods, particularly for those using the ResNet50 network. Again we see similar small regions identified by RISE despite taking a much longer time to generate.

6.3 CUB Birds 200

Figure 7 contains five examples from the CUB Birds dataset. This dataset contains a number of birds that are small in the image compared to ImageNet objects. In this context our methods ability to explain small regions within an image is very useful. For example in Figure 7c for the ResNet50 network, we can see the varying granularity stages of explanation the different methods can produce. The CAM methods produce a heatmap that covers the entire bird, whilst RISE is able to reduce this to only covering the head of the bird. In contrast, our method is able to identify regions within the birds head that are important to the networks predictions.

6.4 Oxford Flowers 102

Figure 8 contains five examples from the Oxford Flowers dataset. The flowers dataset is the opposite of CUB Birds with large objects dominating the frame. In images where there is a single distinctive feature (e.g. the centre of the flower in Figure 8d and e) we see that our method produces a small, compact heatmap around these regions. As with the barbershop image in Figure 5, we see that images such as Figure 8a can produce noisy explanations. However, we also note that a number of the alternative explanations for this image are also noisy.

6 ADDITIONAL QUALITATIVE RESULTS - IMAGES

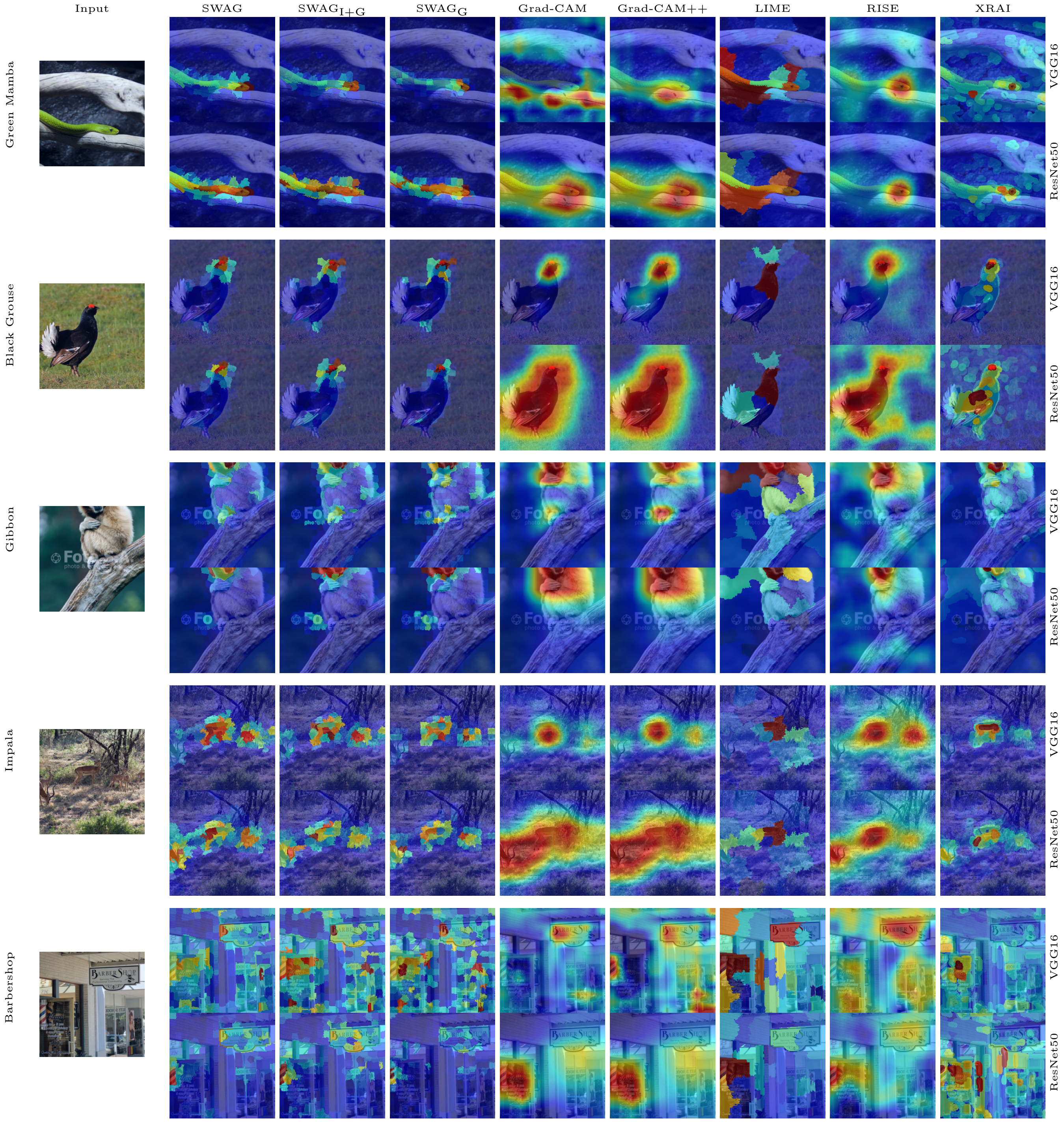

Figure 4: Examples taken from the ImageNet validation set

6 ADDITIONAL QUALITATIVE RESULTS - IMAGES

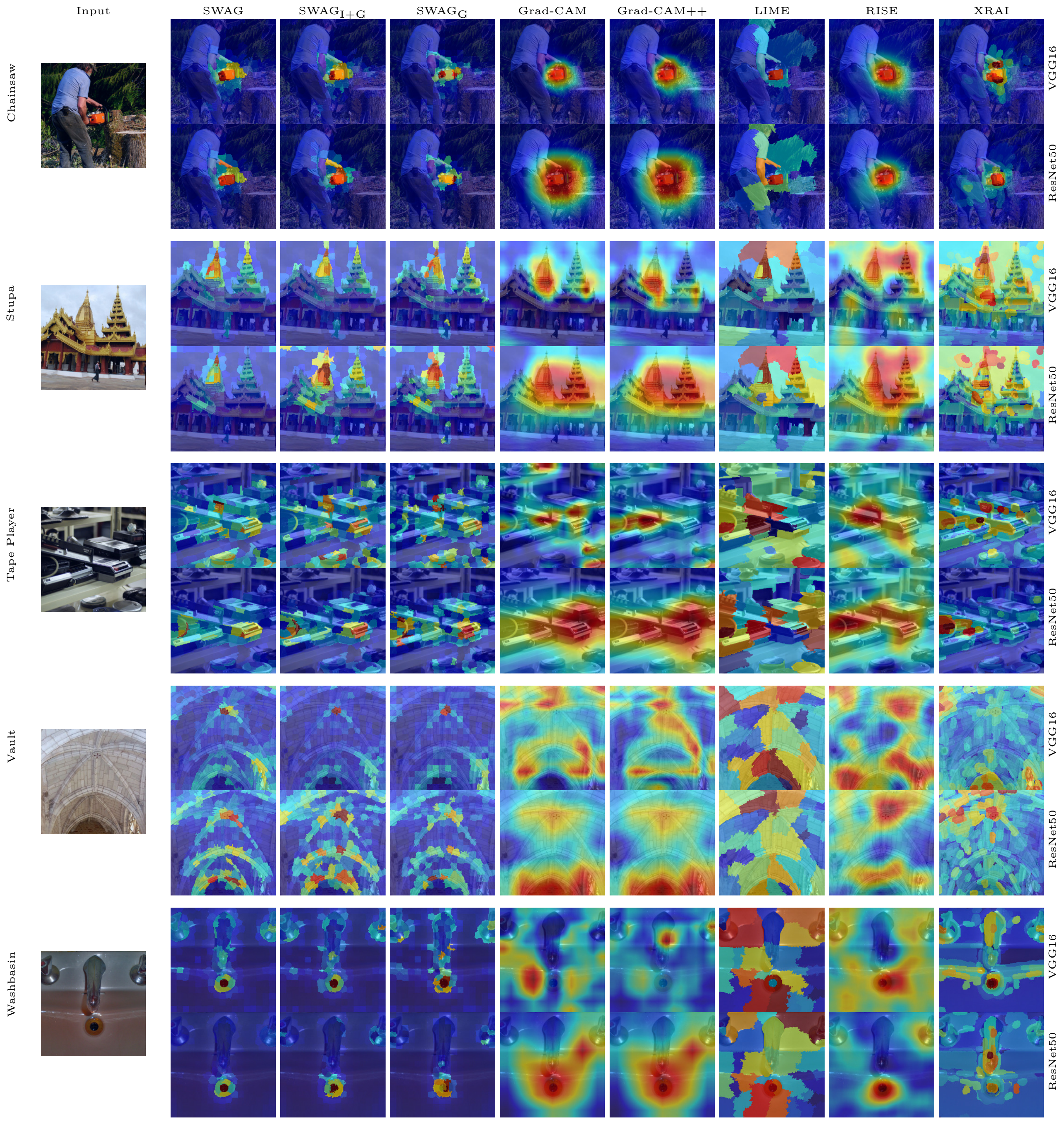

Figure 5: Further examples taken from the ImageNet validation set

6 ADDITIONAL QUALITATIVE RESULTS - IMAGES

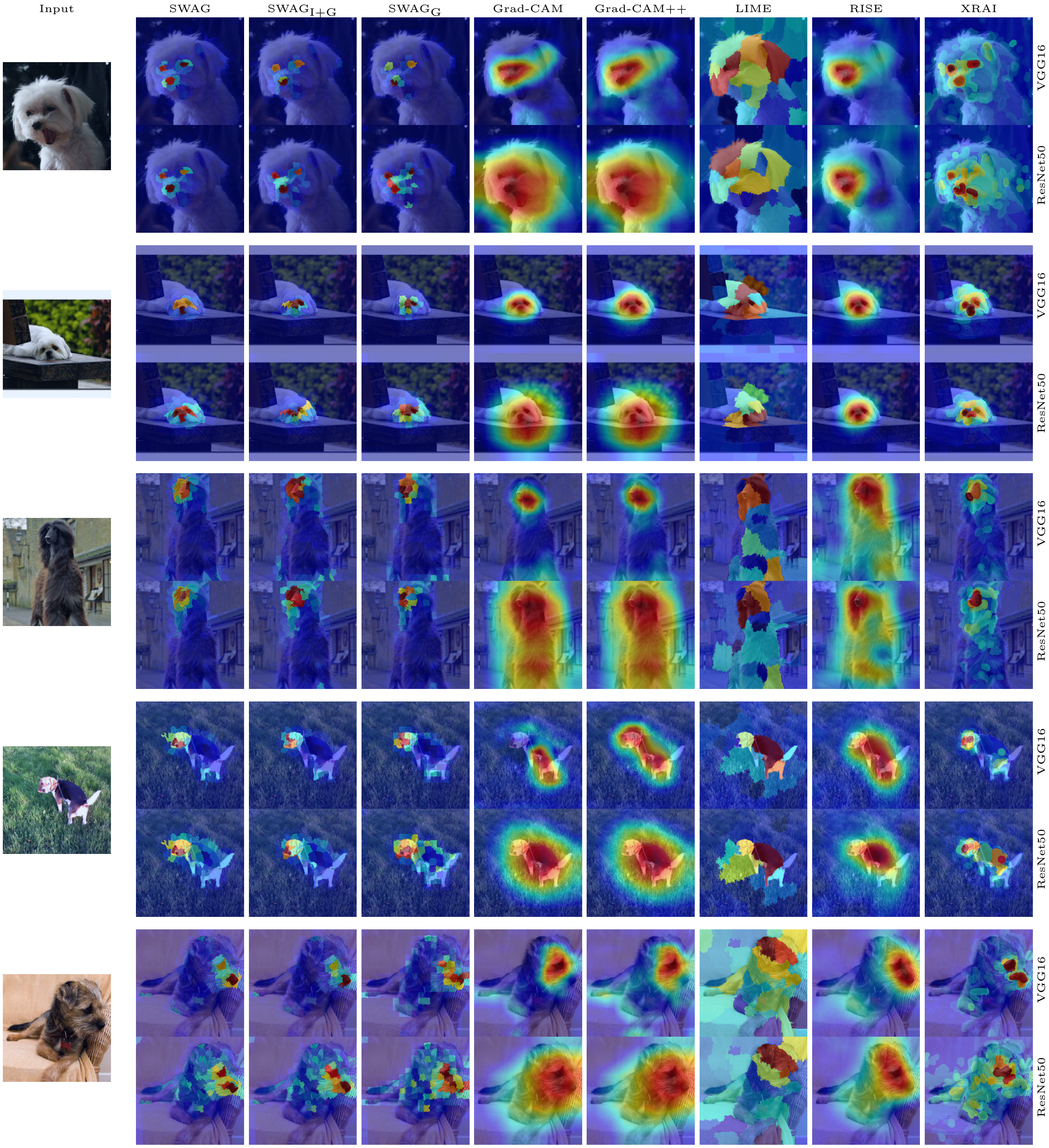

Figure 6: Examples taken from the Stanford Dogs validation set

6 ADDITIONAL QUALITATIVE RESULTS - IMAGES

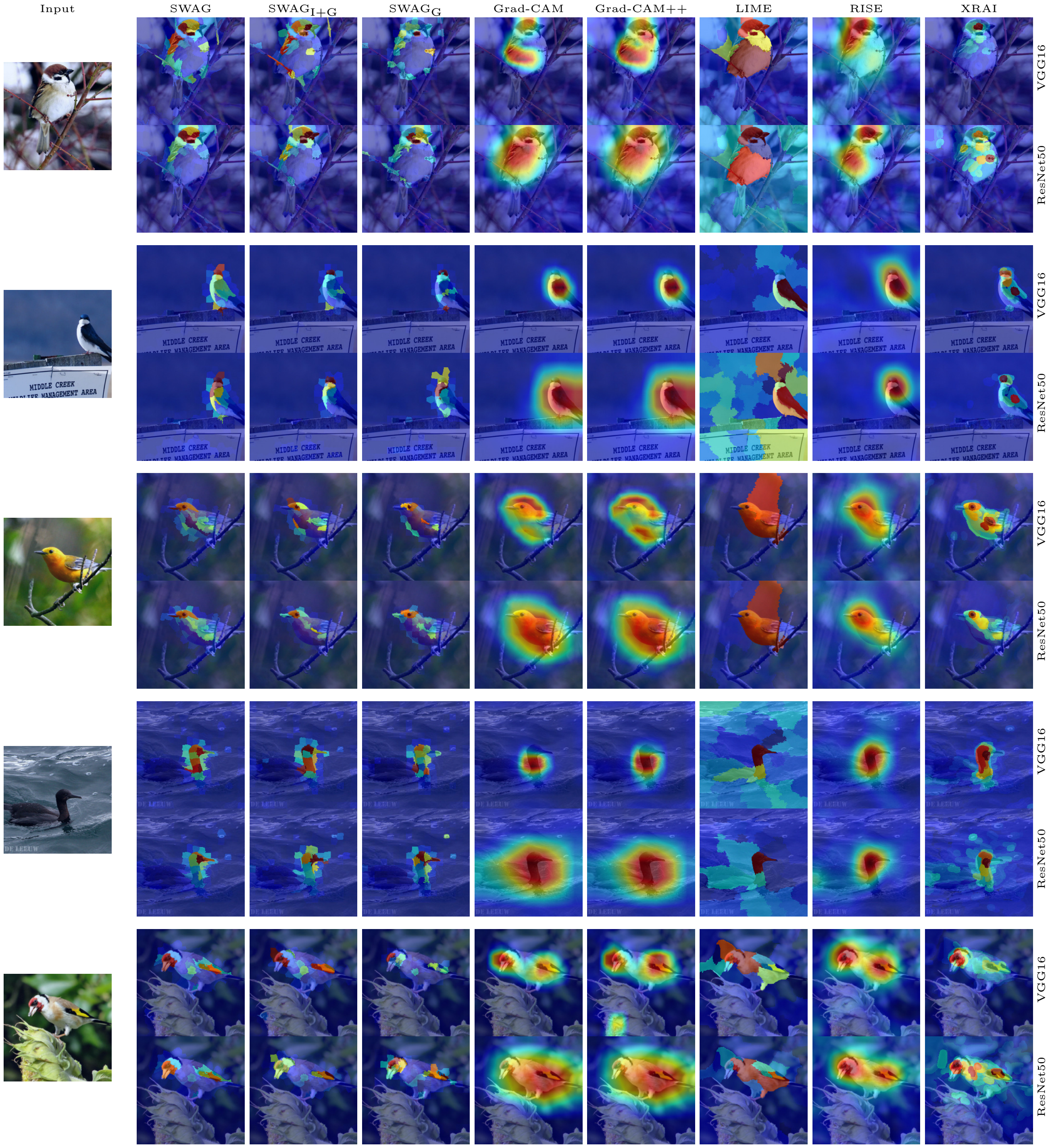

Figure 7: Examples taken from the CUB Birds validation set

6 ADDITIONAL QUALITATIVE RESULTS - IMAGES

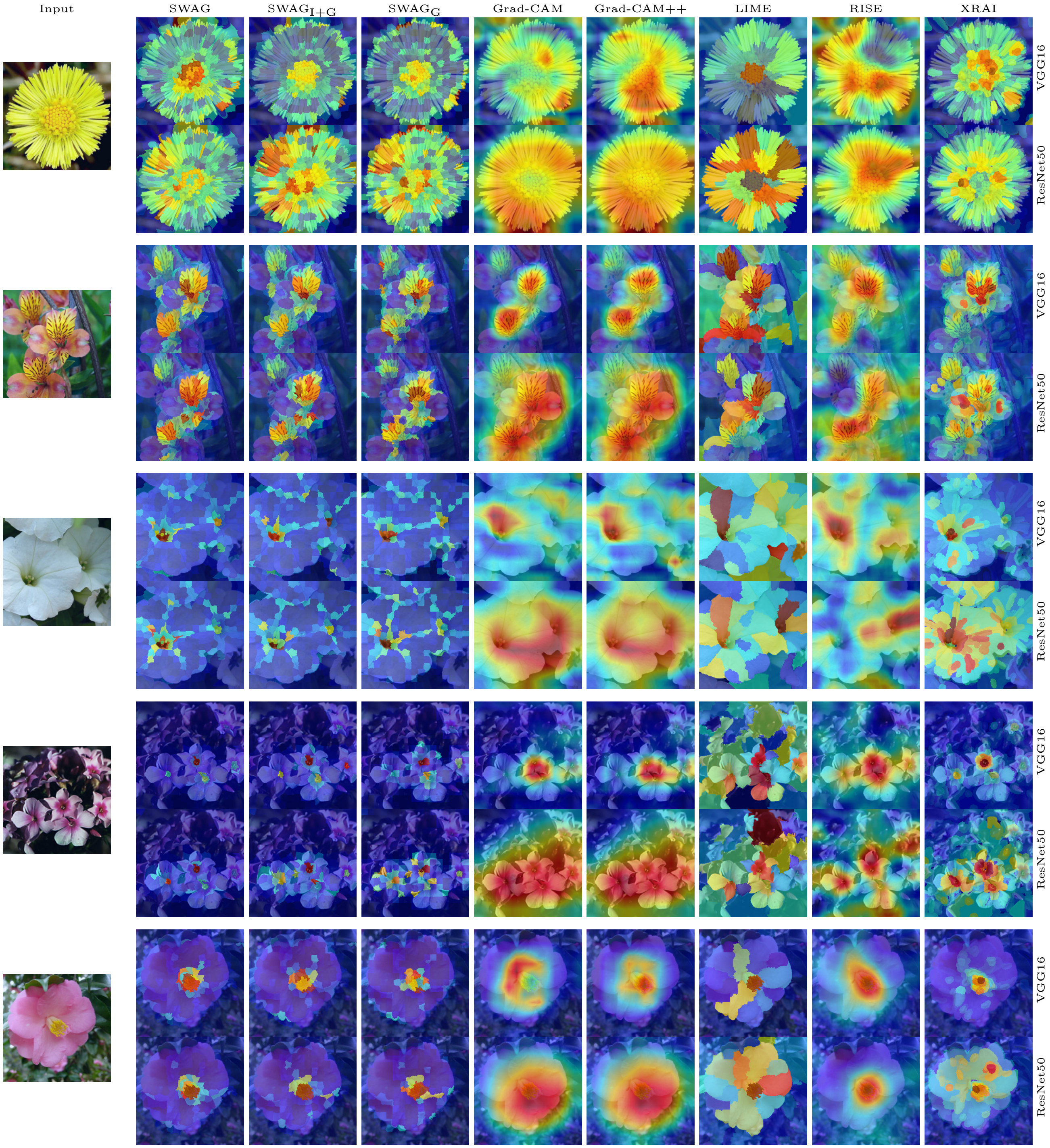

Figure 8: Examples taken from the Oxford Flowers validation set

References

[1] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 839–847, 2018.

[2] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, Oct 2017.

[3] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC*, 2018.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[5] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Oct 2017.