Supplementary Material: Explaining Classifiers using Adversarial Perturbations on the Perceptual Ball

A. Additional ImageNet Examples

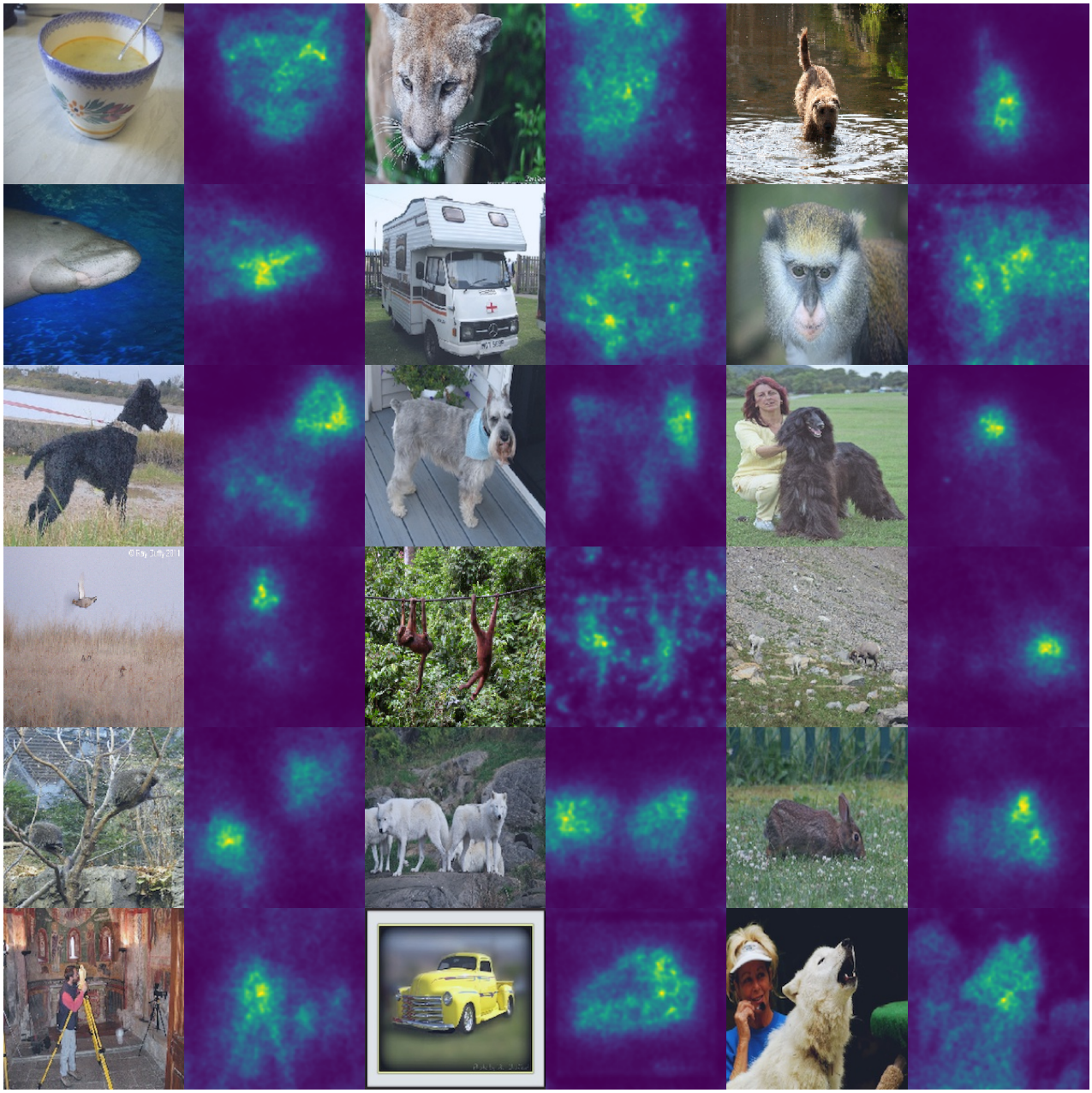

Figure S1. Additional Perceptual Perturbations on ImageNet as explanations. Illustration of perceptual perturbations on typical images taken from ImageNet. The saliency maps do well on localising single object (cougar) which also tends to focus upon the heads of the labelled class. However there are some errors in localisation when supporting classes are also present in the image. For example, see the children with tricycle. While the presence of the child on the tricycle, might help with localisation, the other child is also labeled as salient. See discussion in Section 4 of main paper for more explanation.

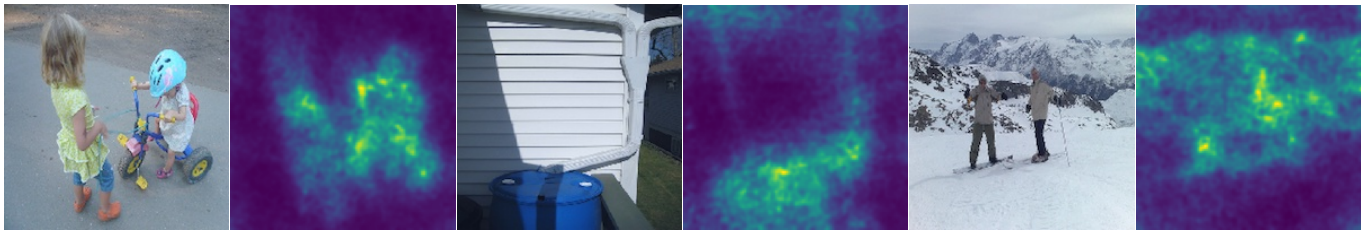

Figure S2. Failed explanation examples. From left to right. (i) the children near the tricycle; (ii) the pipes of the rain barrel; (iii) people in front of the alp; are found to be salient.

B. Additional Results

In this section, we display a small number of additional results to complement the results in the paper. This consists of additional ablation results, for the insertion deletion game (Fig. S3), and the pointing game (Fig. S4). This section also contains a discussion of the parameters in the pointing game with Sec. B.1 discussing the parameter choice for our no perceptual baseline, and Sec. B.2 presenting results on the pointing game with a more limited set of σs.

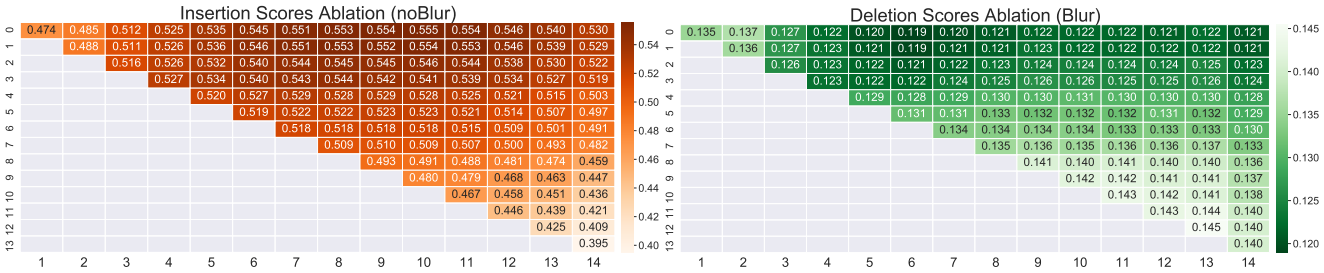

Figure S3. Additional Sensitivity study on the choice of layers for the insertion deletion game. On left we display the Insertion results for the no Blur variant (higher is better) and on the right we display the Deletion results for the blur variant (lower is better).

B.1. Sigma Selection for the No Perceptual baseline in the pointing game

To compare against the no perceptual baseline in the pointing game, we select the best sigma using our ablation study. In the non resized case $\sigma \in \{61.0, 62.0\}$, and we select the largest value 62.0. For the resized case, $\sigma \in \{31.0, 32.0, 33.0, 36.0, 37.0, 49.0, 52.0\}$ have equal performance and we select 37.0, as it is the largest value in the mostly continuous sequence from 31 to 37.

B.2. Pointing Game with Limited σ

The range of σ used in the pointing game is relatively large with 101 possible values. Further, as we are using 0 padded blurs, the larger values of σ will concentrate more on the center of the image.

In this supplementary, we restrict the maximum value of σ to 25, which is in the same range as the value used in Extremal Perturbation (≈ 20) and rerun the experiment. The ablation shown in Fig. S5, shows similar results, with a slight decrease in

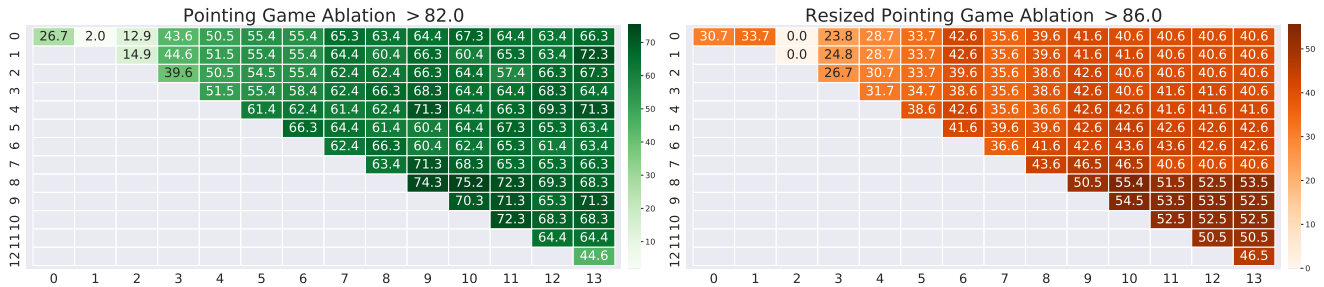

Figure S4. Sensitivity study on the choice of layers regularized for the pointing game. To test how robust each result we display the number of σs which have performance above 82% for the regular image size (left) and 86% for the resized image (right).

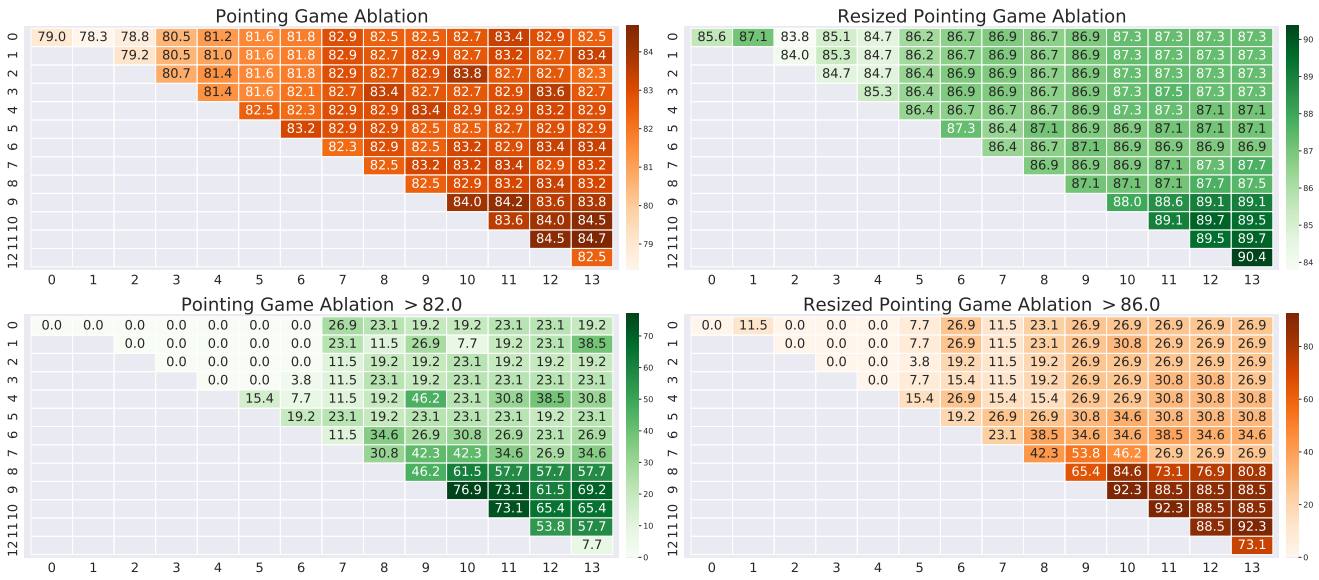

Figure S5. Sensitivity study on the choice of layers regularized for the pointing game with limited σ. To test how robust each result we display the number of σs which have performance above 82% for the regular image size (left) and 86% for the resized image (right).

performance (0.4% in the standard game), which is to be expected as we are maximizing over a smaller number of values of σ.

We select the set of layers and $\sigma \in \{0, \dots 25\}$ which maximizes the performance on the ablation set, which for the regular set is ReLU 11 and 12 with $\sigma = 21$, and for the resized set it is ReLU 12 with $\sigma = 23.0$. We also compare against the no perceptual baseline with the same restricted σ, in both cases a σ of 25 is optimal.

Method	Orig. Image	Scaled Image
Us NoPer	78.8 (58.9)	85.3 (70.5)
Us	84.8 (70.5)	87.7 (74.5)

Table S1. Results for the reduced σ in the pointing game. This table differs from the version in the main text, as for this experiment we further limit $\sigma \in \{0, \dots 25\}$. We present two sets of results, the performance on the images in the original size (first column), and the performance on images which are scaled to 1.5x using bilinear scaling (second column). For each result, the first number shows the percentage on the standard set, with the number in brackets the performance on the difficult set.

The results can be seen in Fig. S5 and Table S1. The results are broadly comparable with the larger σ study. The main change is that the range of layers that obtain over 82% (and over 86% in the resized study) is reduced, although the range of layers that perform well is maintained. For regular sized images there is a 0.3% drop in the standard study and an improvement in the difficult study. The performance in the resized study is comparable with a 0.5% drop in the resized study and a 2.0% drop in the resized difficult study. The no perceptual baseline results have a more substantial drop with the original image size, potentially indicating that the blur is more important for the no perceptual method, in the standard set with a drop of 2.4% and a drop of 4% in the difficult set. Whereas in the resized set, there is a drop of 0.7% and 1.1%. In the limited σ the perceptual outperforms no perceptual in all cases, further, even comparing the limited σ perceptual results with the general σ no perceptual results, the perceptual results still outperforms in all cases. Further, regardless of the σ set, the order and the qualitative performance of the alternative methods in the standard set is maintained.