
Real Time Image Saliency for Black Box Classifiers

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 In this work we develop a fast saliency detection method that can be applied to
2 any differentiable image classifier. We train a masking model to manipulate the
3 scores of the classifier by masking salient parts of the input image. Our model
4 generalises well to unseen images and requires a single forward pass to perform
5 saliency detection, therefore suitable for use in real-time systems. We test our
6 approach on Cifar-10 and ImageNet datasets and show that the produced saliency
7 maps are easily interpretable, sharp, and free of artifacts. We suggest a new metric
8 for saliency and test our method on the ImageNet object localisation task. We
9 achieve results outperforming other weakly supervised methods.

10 1 Introduction

11 The current state of the art image classifiers rival human performance on image classification tasks,
12 but often exhibit unexpected and unintuitive behaviour [7, 14]. For example, we can apply a small
13 perturbation to the input image, unnoticeable to the human eye, to completely fool a classifier [14].

14 Another example of an unexpected behaviour is when a classifier fails to *understand* a given class
15 despite having high accuracy. For example, if “polar bear” is the only class in the dataset that contains
16 snow, a classifier may be able to get a 100% accuracy on this class by simply detecting the presence
17 of snow and ignoring the bear completely [7]. Therefore, even with perfect accuracy, we can not
18 be sure whether our model actually detects polar bears or just snow. One way to decouple the two
19 would be to find snow-only or polar-bear-only images and evaluate the model’s performance on these
20 images separately. An alternative is to use an image of a polar bear with snow from the dataset and
21 apply a *saliency detection method* to test what the classifier is really looking at [7, 12].

22 Saliency detection methods show which parts of a given image are the most relevant to the model
23 for a selected class. Such saliency maps can be obtained for example by finding the smallest region
24 whose removal causes the classification score to drop significantly. This is because we expect the
25 removal of a patch which is not useful for the model not to affect the classification score much.
26 Finding such a salient region can be done iteratively, but this usually requires hundreds of iterations
27 and is therefore a time consuming process.

28 In this paper we lay the groundwork for a new class of fast and accurate model-based saliency
29 detectors, giving high pixel accuracy and sharp saliency maps (an example is given in figure 1). We
30 propose a fast, model agnostic, saliency detection method. Instead of iteratively obtaining saliency
31 maps for each input image separately, we train a model to predict such a map for any input image in a
32 single feed-forward pass. We show that this approach is not only orders-of-magnitude faster than
33 iterative methods, but it also produces higher quality saliency masks and achieves better localisation
34 results. We assess this with standard saliency benchmarks, and introduce a new saliency measure.
35 Our proposed model is able to produce real-time saliency maps, enabling new saliency applications
36 such as video-saliency which we comment on in our *Future Research* section.

37 2 Related work

38 Since the rise of CNNs in 2012 [6] numerous methods of image saliency detection have been proposed.
39 One of the earliest such methods is a gradient-based approach introduced in [12] which computes
40 the gradient of the class with respect to the image and assumes that salient regions are at locations

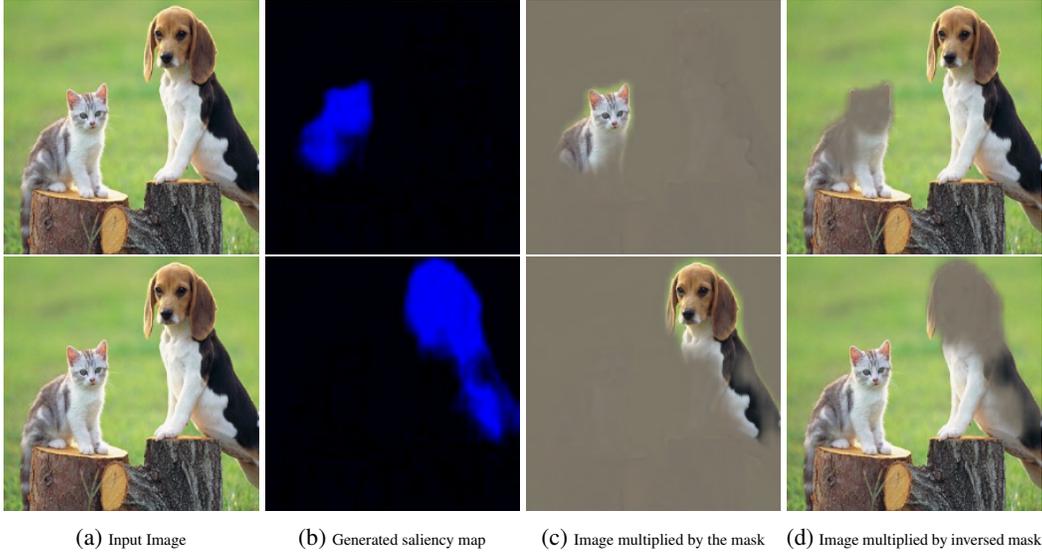


Figure 1: An example of explanations produced by our model. The top row shows the explanation for the "Egyptian cat" while the bottom row shows the explanation for the "Beagle". Note that produced explanations can precisely both highlight and remove the selected object from the image.

41 with high gradient magnitude. Other similar backpropagation-based approaches have been proposed,
 42 for example Guided Backpropagation [13] or Excitation Backprop [17]. While the gradient based
 43 methods are fast enough to be applied in real-time, they produce explanations of limited quality [17]
 44 and they are hard to improve and build upon.

45 Zhou et al. [18] proposed an approach that iteratively removes patches of the input image (by setting
 46 them to the mean colour) such that the class score is preserved. After a sufficient number of iterations,
 47 we are left with salient parts of the original image. The maps produced by this method are easily
 48 interpretable, but unfortunately, the iterative process is very time consuming and not acceptable for
 49 real-time saliency detection.

50 In another work, Cao et al. [2] introduced an optimisation method that aims to preserve only a fraction
 51 of network activations such that the class score is maximised. Again, after the iterative optimisation
 52 process, only activations that are relevant remain and their spatial location in the CNN feature map
 53 indicate salient image regions.

54 Very recently (and in parallel to this work), another optimisation based method was proposed [3].
 55 Similarly to Cao et al. [2], Fong and Vedaldi [3] also propose to use gradient descent to optimise for
 56 the salient region, but the optimisation is done only in the image space and the classifier model is
 57 treated as a black box. Essentially Fong and Vedaldi [3]’s method tries to remove as little from the
 58 image as possible, and at the same time to reduce the class score as much as possible. A removed
 59 region is then a minimally salient part of the image. This approach is model agnostic and the produced
 60 maps are easily interpretable because the optimisation is done in the image space and the model is
 61 treated as a black box.

62 We next argue what conditions a good saliency model should satisfy, and propose a new metric for
 63 saliency.

64 3 Image Saliency and Introduced Evidence

65 Image saliency is relatively hard to define and there is no single obvious metric that could measure
 66 the quality of the produced map. In simple terms, the saliency map is defined as a summarised
 67 explanation of where the classifier “looks” at to make its prediction.

68 There are two slightly more formal definitions of saliency that we can use:

- 69 • Smallest sufficient region (SSR) — smallest region of the image that alone allows a confident
 70 classification,

- Smallest destroying region (SDR) — smallest region of the image that when removed, prevents a confident classification.

Similar concepts were suggested in [3] as well. An example of SSR and SDR is shown in figure 2. It can be seen that SSR is very small and has only one seal visible. Given this SSR, even a human would find it difficult to recognise the preserved image. Nevertheless, it contains some characteristic for “seal” features such as parts of the face with whiskers, and the classifier is over 90% confident that this image should be labeled as a “seal”. On the other hand, SDR has a much stronger and larger region and quite successfully removes all the evidence for seals from the image. In order to be as informative as possible we would like to find a region that performs well as both SSR and SDR.

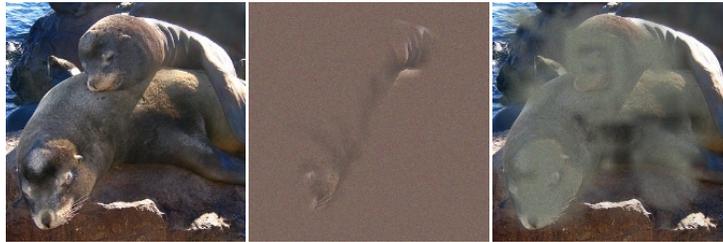


Figure 2: From left to right: the input image; smallest sufficient region (SSR); smallest destroying region (SDR). Regions were found using the mask optimisation procedure from [3].

Both SDR and SSR remove some evidence from the image. There are few ways of removing evidence, for example by blurring the evidence, setting it to a constant colour, adding noise, or by completely cropping out the unwanted parts. Unfortunately, each one of these methods introduces new evidence that can be used by the classifier as a side effect. For example, if we remove a part of the image by setting it to the constant colour green then we may also unintentionally provide evidence for “grass” which in turn may increase the probability of classes appearing often with grass (such as “giraffe”). We discuss this problem and ways of minimising introduced evidence next.

3.1 Fighting the Introduced Evidence

As mentioned in the previous section, by manipulating the image we always introduce some extra evidence. Here, let us focus on the case of applying a mask M to the image X to obtain the edited image E . In the simplest case we can simply multiply X and M element-wise:

$$E = X \odot M \tag{1}$$

This operation sets certain regions of the image to a constant “0” colour. While setting a larger patch of the image to “0” may sound rather harmless (perhaps following the assumption that the mean of all colors carries very little evidence), we may encounter problems when the mask M is not *smooth*. The mask M , in the worst case, can be used to introduce a large amount of additional evidence by generating adversarial artifacts (a similar observation was made in [3]). An example of such a mask is presented in figure 3. Adversarial artifacts generated by the mask are very small in magnitude and almost imperceptible for humans, but they are able to completely destroy the original prediction of the classifier. Such adversarial masks provide very poor saliency explanations and therefore should be avoided.

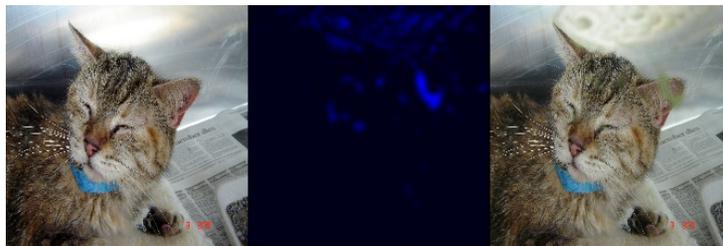


Figure 3: The adversarial mask introduces very small perturbations, but can completely alter the classifier’s predictions. From left to right: an image which is correctly recognised by the classifier with a high confidence as a “tabby cat”; a generated adversarial mask; an original image after application of the mask that is no longer recognised as a “tabby cat”.

100 There are a few ways to make the introduction of artifacts harder. For example, we may change the
 101 way we apply a mask to reduce the amount of unwanted evidence due to specifically-crafted masks:

$$E = X \odot M + A \odot (1 - M) \quad (2)$$

102 where A is an alternative image. A can be chosen to be for example a highly blurred version of X .
 103 In such case mask M simply selectively adds blur to the image X and therefore it is much harder
 104 to generate high-frequency-high-evidence artifacts. Unfortunately, applying blur does not eliminate
 105 existing evidence very well, especially in the case of images with low spatial frequencies like a
 106 seashore or mountains.

107 Another reasonable choice of A is a random constant colour combined with high-frequency noise.
 108 This makes the resulting image E more unpredictable at regions where M is low and therefore it is
 109 slightly harder to produce a reliable artifact.

110 Even with all these measures, adversarial artifacts may still occur and therefore it is necessary to
 111 encourage smoothness of the mask M for example via a total variation (TV) penalty. We can also
 112 directly resize smaller masks to the required size as resizing can be seen as a smoothness mechanism.

113 3.2 A New Saliency Metric

114 To assess the quality and interpretability of saliency maps we introduce a new saliency metric.
 115 According to the SSR objective we require that the classifier is able to still recognise the object from
 116 the preserved region and that the preserved region is as small as possible. In order to make sure that
 117 the preserved region is free from adversarial artifacts, instead of masking we can crop the image. We
 118 propose to find the tightest rectangular crop that *contains the entire salient region* and to feed that
 119 rectangular region to the classifier to directly verify whether it is able to recognise the requested class.
 120 We define our saliency metric simply as:

$$s(a, p) = \log(\tilde{a}) - \log(p) \quad (3)$$

121 with $\tilde{a} = \max(a, 0.05)$. Here a is the area of the rectangular crop as a fraction of the total image size
 122 and p is the probability of the requested class returned by the classifier based on the cropped region.
 123 The metric is almost a direct translation of the SSR. We threshold the area at 0.05 in order to prevent
 124 instabilities at low area fractions. Good saliency detectors will be able to significantly reduce the
 125 crop size without reducing the classification probability, and therefore a low value for the saliency
 126 metric is a characteristic of good saliency detectors.

127 Interpreting this metric following *information theory*, this measure can be seen as the relative amount
 128 of information between an indicator variable with probability p and an indicator variable with
 129 probability a —or the *concentration of information* in the cropped region.

130 Because most image classifiers accept only images of a fixed size and the crop can have an arbitrary
 131 size, we resize the crop to the required size disregarding aspect ratio. This seems work well in
 132 practice.

133 3.3 The Saliency Objective

134 Taking the previous conditions into consideration we want to find a mask M that is smooth and
 135 performs well at both SSR and SDR; examples of such masks can be seen in figure 1. Therefore,
 136 more formally, given class c of interest, and an input image X , to find a saliency map M for class c ,
 137 our objective function L is given by:

$$L(M) = \lambda_1 TV(M) + \lambda_2 AV(M) - \log(f_c(\Phi(X, M))) + \lambda_3 f_c(\Phi(X, 1 - M))^{\lambda_4} \quad (4)$$

138 where f_c is a softmax probability of the class c of the black box image classifier and $TV(M)$ is the
 139 total variation of the mask defined simply as:

$$TV(M) = \sum_{i,j} (M_{ij} - M_{ij+1})^2 + \sum_{i,j} (M_{ij} - M_{i+1j})^2, \quad (5)$$

140 $AV(M)$ is the average of the mask elements, taking value between 0 and 1, and λ_i are regularisers.
 141 Finally, the function Φ removes the evidence from the image as introduced in the previous section:

$$\Phi(X, M) = X \odot M + A \odot (1 - M). \quad (6)$$

142 In total, the objective function is composed of 4 terms. The first term enforces mask smoothness,
 143 the second term encourages that the region is small. The third term makes sure that the classifier is

144 able to recognise the selected class from the preserved region. Finally, the last term ensures that the
 145 probability of the selected class, after the salient region is removed, is low (note that the inverted
 146 mask $1 - M$ is applied). Setting λ_4 to a value smaller than 1 (e.g. 0.2) helps reduce this probability
 147 to very small values.

148 4 Masking Model

149 The mask can be found iteratively for a given image-class pair by directly optimising the objective
 150 function from equation 4. In fact, this is the method used for the by [3] which was developed in
 151 parallel to this work, with the only difference that [3] only optimise the mask iteratively and for SDR
 152 (so they don't include the third term of our objective function). Unfortunately, iteratively finding the
 153 mask is not only very slow, as normally more than 100 iterations are required, but it also causes the
 154 mask to greatly overfit to the image and a large TV penalty is needed to prevent adversarial artifacts
 155 from forming. Therefore, the produced masks are blurry, imprecise, and overfit to the specific image
 156 rather than capturing the general behaviour of the classifier (see figure 2).

157 For the above reasons, we develop a trainable masking model that can produce the desired masks
 158 in a single forward pass without direct access to the image classifier after training. The masking
 159 model receives an image and a class selector as inputs and learns to produce masks that minimise our
 160 objective function (equation 4). In order to succeed at this task, the model must learn which parts
 161 of the input image are considered salient by the black box classifier. In theory, the model can still
 162 learn to develop adversarial masks that perform well on the objective function, but in practice it is
 163 not an easy task, because adversarial patterns for all the classes would have to be learned in order to
 164 perform well on the SSR part of the objective.

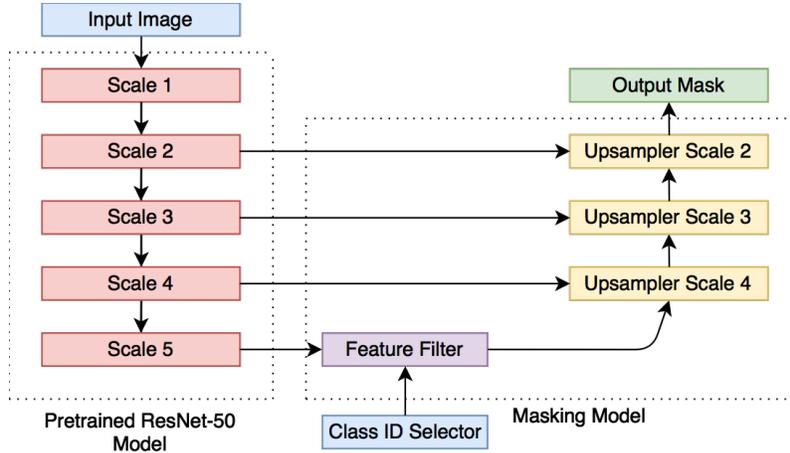


Figure 4: Architecture diagram of the masking model.

165 In order to make our masks sharp and precise, we adapt a U-Net architecture [9] so that the masking
 166 model can use feature maps from multiple resolutions. The architecture diagram can be seen in figure
 167 4. For the encoder part of the U-Net we use ResNet-50 [4] pre-trained on ImageNet [10].

168 The ResNet-50 model contains feature maps of five different scales, where each subsequent scale
 169 block downsamples the input by a factor of two. We use the ResNet's feature map from Scale 5
 170 (which corresponds to downsampling by a factor of 32) and pass it through the feature filter. The
 171 purpose of the feature filter is to attenuate spatial locations which contents do not correspond to
 172 the selected class. Therefore, the feature filter performs the initial localisation, while the following
 173 upsampling blocks fine-tune the produced masks. The output of the feature filter Y at spatial location
 174 i, j is given by:

$$Y_{ij} = X_{ij} \sigma(X_{ij}^T C_s) \quad (7)$$

175 where X_{ij} is the output of the Scale 5 block at spatial location i, j ; C_s is the embedding of the
 176 selected class s and $\sigma(\cdot)$ is the sigmoid nonlinearity. Class embedding C can be learned as part of the
 177 overall objective.

178 The upsampler blocks take the lower resolution feature map as input and upsample it by a factor
 179 of two using transposed convolution [16], afterwards they concatenate the upsampled map with the
 180 corresponding feature map from ResNet and follow that with three bottleneck blocks [4].

181 Finally, to the output of the last upsampler block (Upsampler Scale 2) we apply 1x1 convolution to
182 produce a feature map with with just two channels — C_0, C_1 . The mask M_s is obtained from:

$$M_s = \frac{\text{abs}(C_0)}{\text{abs}(C_0) + \text{abs}(C_1)} \quad (8)$$

183 We use this nonstandard nonlinearity because sigmoid and tanh nonlinearities did not optimise
184 properly and the extra degree of freedom from two channels greatly improved training. The mask M_s
185 has resolution four times lower than the input image and has to be upsampled by a factor of four with
186 bilinear resize to obtain the final mask M .

187 The complexity of the model is comparable to that of ResNet-50 and it can process more than a
188 hundred 224x224 images per second on a standard GPU (which is sufficient for real time saliency
189 detection).

190 4.1 Training process

191 We train the masking model to directly minimise the objective function from equation 4. The weights
192 of the pre-trained ResNet encoder (red blocks in figure 4) are kept fixed during the training.

193 In order to make the training process work properly, we introduce few optimisations. First of all,
194 in the naive training process the ground truth label would always be supplied as a class selector.
195 Unfortunately, under such setting, the model learns to completely ignore the class selector and simply
196 always masks the dominant object in the image. The solution to this problem is to sometimes supply
197 a class selector for a fake class and to apply only the area penalty term of the objective function.
198 Under this setting the model must pay attention to the class selector, as the only way it can reduce
199 loss in case of a fake label is by setting the mask to zero. During training, we set the probability of
200 the fake label occurrence to 30%. One can also greatly speed up the embedding training by ensuring
201 that the maximal value of $\sigma(X_{ij}^T C_s)$ from equation 7 is high in case of a correct label and low in case
202 of a fake label.

203 Finally, let us consider again the evidence removal function $\Phi(X, M)$. In order to prevent the model
204 from adapting to any single evidence removal scheme the alternative image A is randomly generated
205 every time the function Φ is called. In 50% of cases the image A is the blurred version of X (we use
206 a Gaussian blur with $\sigma = 10$ to achieve a strong blur) and in the remainder of cases, A is set to a
207 random colour image with the addition of a Gaussian noise. Such a random scheme greatly improves
208 the quality of the produced masks as the model can no longer make strong assumptions about the
209 final look of the image.

210 5 Experiments

211 In the ImageNet saliency detection experiment we use three different black-box classifiers: AlexNet
212 [6], GoogleNet [15] and ResNet-50 [4]. These models are treated as black boxes and for each one
213 we train a separate masking model. The selected parameters of the objective function are $\lambda_1 = 10$,
214 $\lambda_2 = 10^{-3}$, $\lambda_3 = 5$, $\lambda_4 = 0.3$. The first upsampling block has 768 output channels and with each
215 subsequent upsampling block we reduce the number of channels by a factor of two. We train each
216 masking model as described in section 4.1 on 250,000 images from the ImageNet training set. During
217 the training process, a very meaningful class embedding was learned and we include its visualisation
218 in the Appendix.

219 Example masks generated by the saliency models trained on three different black box image classifiers
220 can be seen in figure 5, where the model is tasked to produce a saliency map for the ground truth
221 label. In figure 5 it can be clearly seen that the quality of masks generated by our models clearly
222 outperforms alternative approaches. The masks produced by models trained on GoogleNet and
223 ResNet are sharp and precise and would produce accurate object segmentations. The saliency model
224 trained on AlexNet produces much stronger and slightly larger saliency regions, possibly because
225 AlexNet is a less powerful model which needs more evidence for successful classification.

226 5.1 Weakly supervised object localisation

227 A possible metric to evaluate saliency maps is by object localisation. We adapt the evaluation protocol
228 from [2] and provide the ground truth label to the masking model. Afterwards, we threshold the
229 produced saliency map at 0.5 and the tightest bounding box that contains the whole saliency map is

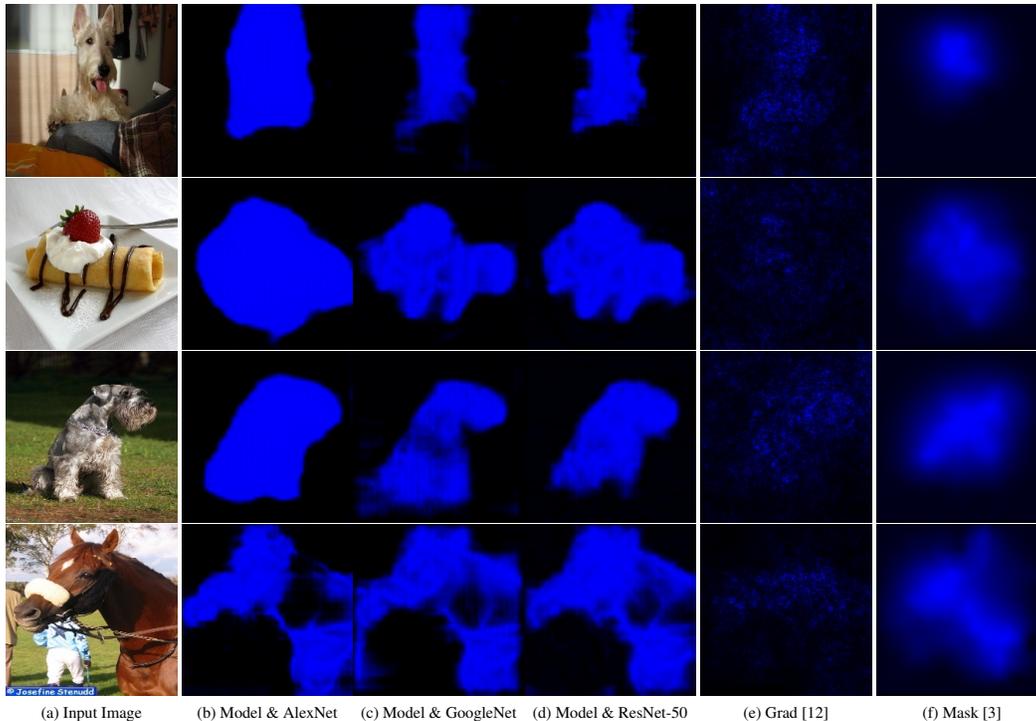


Figure 5: Saliency maps generated by different methods for the ground truth class. The ground truth classes, starting from the first row are: Scottish terrier, chocolate syrup, standard schnauzer and sorrel. Columns b, c, d show the masks generated by *our* masking models, each trained on a different black box classifier (from left to right: AlexNet, GoogleNet, ResNet-50). Last two columns e, f show saliency maps for GoogleNet generated respectively by gradient [12] and the recently introduced iterative mask optimisation approach [3].

230 set as the final localisation box. The localisation box has to have IOU greater than 0.5 with any of the
 231 ground truth bounding boxes in order to consider the localisation successful, otherwise, it is counted
 232 as an error. The calculated error rates for the three models are presented in table 1. The lowest
 233 localisation error of 36.7% was achieved by the saliency model trained on the ResNet-50 black box,
 234 this is a good achievement considering the fact that our method was not given any localisation training
 235 data and that a fully supervised approach employed by VGG [11] achieved only slightly lower error
 236 of 34.3%. The localisation error of the model trained on GoogleNet is very similar to the one trained
 237 on ResNet. This is not surprising because both models produce very similar saliency masks (see
 238 figure 5). The AlexNet trained model, on the other hand, has a considerably higher localisation error
 239 which is probably a result of AlexNet needing larger image contexts to make a successful prediction
 240 (and therefore producing saliency masks which are slightly less precise).

	Alexnet [6]	GoogleNet [15]	ResNet-50 [4]
Localisation Err (%)	39.8	36.9	36.7

Table 1: Weakly supervised bounding box localisation error on ImageNet validation set for our masking models trained with different black box classifiers.

241 We also compared our object localisation errors to errors achieved by other weakly supervised
 242 methods and existing saliency detection techniques. As a baseline we calculated the localisation error
 243 of the centrally placed rectangle which spans half of the image area — which we name "Center".
 244 The results are presented in table 2. It can be seen that our model outperforms other approaches,
 245 sometimes by a significant margin. It also performs significantly better than the baseline (centrally
 246 placed box) and the iteratively optimised saliency masks. Because a big fraction of ImageNet images
 247 have a large, dominant object in the center, the localisation accuracy of the centrally placed box is
 248 relatively high and it managed to outperform two methods from previous literature.

Center	Grad [12]	Guid [13]	LRP [1]	CAM [19]	Exc [17]	Feed [2]	Mask [3]	This Work
46.3	41.7	42.0	57.8	48.1	39.0	38.7	43.1	36.9

Table 2: Localisation errors(%) on ImageNet validation set for popular weakly supervised methods. Error rates were taken from [3] which recalculated originally reported results using few different mask thresholding techniques and achieved slightly lower error rates. For a fair comparison, all the methods follow the same evaluation protocol of [2] and produce saliency maps for GoogleNet classifier [15].

249 5.2 Evaluating the saliency metric

250 To better assess the interpretability of the produced masks we calculate the saliency metric introduced
 251 in section 3.2 for selected saliency methods and present the results in the table 3. We include a few
 252 baseline approaches — the "Central box" introduced in the previous section, and the "Max box"
 253 which simply corresponds to a box spanning the whole image. We also calculate the saliency metric
 254 for the ground truth bounding boxes supplied with the data, and in case the image contains more than
 255 one ground truth box the saliency metric is set as the average over all the boxes.

256 Table 3 shows that our model achieves a considerably better saliency metric than other saliency
 257 approaches. It also significantly outperforms max box and center box baselines and is on par
 258 with ground truth boxes which supports the claim that the interpretability of the localisation boxes
 259 generated by our model is similar to that of the ground truth boxes.

	Localisation Err (%)	Saliency Metric
Ground truth boxes (baseline)	0.00	0.284
Max box (baseline)	59.7	1.366
Center box (baseline)	46.3	0.645
Grad [12]	41.7	0.451
Exc [17]	39.0	0.415
Masking model (this work)	36.9	0.318

Table 3: ImageNet localisation error and the saliency metric for GoogleNet.

260 Further experiments on Cifar10 are given in the appendix.

261 6 Conclusion and Future Research

262 In this work we have presented a new, fast, and accurate saliency detection method that can be
 263 applied to any differentiable image classifier. Our model is able to produce 100 saliency masks per
 264 second, sufficient for real-time applications. We have shown that our method outperforms other
 265 weakly supervised techniques at the ImageNet localisation task. We have also developed a new
 266 saliency metric that can be used to assess the quality of explanations produced by saliency detectors.
 267 Under this new metric, the quality of explanations produced by our model outperforms other popular
 268 saliency detectors and is on par with ground truth bounding boxes.

269 The model-based nature of our technique means that our work can be extended and built upon by
 270 improving the architecture of the masking network or by changing the objective function to achieve
 271 any desired properties for the output mask.

272 Future work includes modifying the approach to produce high quality, weakly supervised, image
 273 segmentations. More over, because our model can be run in real-time it can be used for video saliency
 274 detection to instantly explain decisions made by black-box classifiers. Lastly, our model might have
 275 biases of its own — a fact which does not seem to influence the model performance in finding biases
 276 in other black boxes according to the various metrics we used. It would be interesting to study the
 277 biases embedded into our masking model itself, and see how these affect the generated saliency
 278 masks.

279 **References**

- 280 [1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and
 281 Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance
 282 propagation. *PLoS ONE*, 10(7):e0130140, 2015. doi: 10.1371/journal.pone.0130140. URL [http://](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4498753/)
 283 www.ncbi.nlm.nih.gov/pmc/articles/PMC4498753/.
- 284 [2] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang,
 285 Chang Huang, Wei Xu, Deva Ramanan, and Thomas S. Huang. Look and think twice: Capturing top-down
 286 visual attention with feedback convolutional neural networks. In *2015 IEEE International Conference*
 287 *on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2956–2964, 2015. doi:
 288 10.1109/ICCV.2015.338. URL <http://dx.doi.org/10.1109/ICCV.2015.338>.
- 289 [3] Ruth Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation,
 290 April 2017. URL <http://arxiv.org/abs/1704.03296>.
- 291 [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition.
 292 *CoRR*, abs/1512.03385, 2015. URL <http://arxiv.org/abs/1512.03385>.
- 293 [5] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Master’s thesis, 2009. URL
 294 <http://www.cs.toronto.edu/~{}kriz/learning-features-2009-TR.pdf>.
- 295 [6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with
 296 deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and
 297 K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages
 298 1097–1105. Curran Associates, Inc., 2012. URL [http://papers.nips.cc/paper/](http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf)
 299 [4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf](http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf).
- 300 [7] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining
 301 the Predictions of Any Classifier. *CoRR*, abs/1602.04938, 2016. URL [http://arxiv.org/abs/1602.](http://arxiv.org/abs/1602.04938)
 302 [04938](http://arxiv.org/abs/1602.04938).
- 303 [8] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua
 304 Bengio. FitNets: Hints for Thin Deep Nets. *CoRR*, abs/1412.6550, 2014. URL [http://arxiv.org/](http://arxiv.org/abs/1412.6550)
 305 [abs/1412.6550](http://arxiv.org/abs/1412.6550).
- 306 [9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical
 307 image segmentation. *CoRR*, abs/1505.04597, 2015. URL <http://arxiv.org/abs/1505.04597>.
- 308 [10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang,
 309 Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large
 310 Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252,
 311 2015. doi: 10.1007/s11263-015-0816-y.
- 312 [11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recogni-
 313 tion. *CoRR*, abs/1409.1556, 2014. URL <http://arxiv.org/abs/1409.1556>.
- 314 [12] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising
 315 image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013. URL [http://arxiv.org/](http://arxiv.org/abs/1312.6034)
 316 [abs/1312.6034](http://arxiv.org/abs/1312.6034).
- 317 [13] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for
 318 simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014. URL [http://arxiv.org/abs/1412.](http://arxiv.org/abs/1412.6806)
 319 [6806](http://arxiv.org/abs/1412.6806).
- 320 [14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow,
 321 and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. URL [http://](http://arxiv.org/abs/1312.6199)
 322 arxiv.org/abs/1312.6199.
- 323 [15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Du-
 324 mitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*,
 325 abs/1409.4842, 2014. URL <http://arxiv.org/abs/1409.4842>.
- 326 [16] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *CoRR*,
 327 abs/1311.2901, 2013. URL <http://arxiv.org/abs/1311.2901>.
- 328 [17] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by
 329 excitation backprop. 2016. URL [https://www.robots.ox.ac.uk/~vgg/rg/papers/zhang_eccv16.](https://www.robots.ox.ac.uk/~vgg/rg/papers/zhang_eccv16.pdf)
 330 [pdf](https://www.robots.ox.ac.uk/~vgg/rg/papers/zhang_eccv16.pdf).
- 331 [18] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Object Detectors Emerge
 332 in Deep Scene CNNs. *CoRR*, abs/1412.6856, 2014. URL <http://arxiv.org/abs/1412.6856>.

333 [19] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for
334 discriminative localization. *CoRR*, abs/1512.04150, 2015. URL <http://arxiv.org/abs/1512.04150>.