We thank the reviewers for their comments.

Reply to #1: While we fully acknowledge that the discussion on the relation to Ghorbani et al. has not made this sufficiently clear, we disagree with the concerns about originality:

We present a method which can manipulate an image to obtain an *arbitrary* target explanation. The methods proposed by Ghorbani et al, including their "targeted" method, cannot structurally reproduce a heatmap. It can only increase the accumulated relevance in a certain subsection of the image. As a result, it is not capable to manipulate the heatmap to closely reproduce a target explanation except for very selected and simple cases. We demonstrate this below. Fine-grained control of the heatmap is absolutely essential for attacks on explanations. We also note that their attacks only keep the classification result the same. This leads to significant changes in the network output (see plot d). From their analysis, it is therefore not clear whether the explanation *or* the network is vulnerable (and the heatmap simply reflects the relevance of the perturbation faithfully). Our method keeps the output constant which is crucial for the geometrical interpretation in terms of principal curvatures and all results derived from it, i.e. all of Sec 3+4.

The final ms will contain a careful discussion on the relation to Ghorbani et al. and a substantially streamlined Section 2. Fig. 3 is moved to the SI. Also we will extend Sec 4 by a discussion of the large scale analysis of β-smoothing, previously in the SI, and additional pixel flipping results (Samek et al 2017, IEEE) establishing that β-smoothing performs better than unsmoothed methods (see plot for a preview).

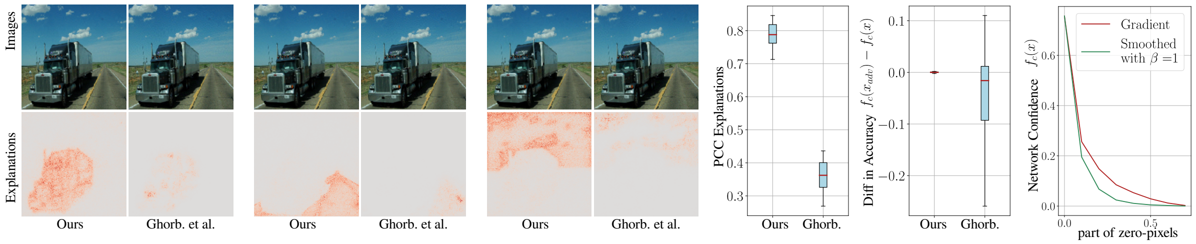
Reply to questions:

- Relevance for relu networks is strongly suggested by relation to SG and indeed confirmed by pixel flipping.
- Beta smoothing increases pixel flipping performance.
- We did not retrain with softplus. We think it is preferable to modify the explanation method since it is less costly.
- In Fig 4, a manipulation of the unsmoothed method is "undone" by smoothing. In all other figures, the smoothed method is attacked directly.
- If both networks have softplus non-linearities, we can compare their bound (9). Note that its constant C depends on the weights of the network.

Summary: Our paper introduces a novel method allowing total control over the heatmap, it explains this manipulability in terms of differential geometry and uses these insights to propose an effective defense with theoretical guarantees. None of these results were contained in Ghorbani et al. We therefore strongly insist on the originality of our paper with respect to Ghorbani et al.

Reply to #3: We completely agree that our algorithm is for differential explainers (remark will be added). The final ms will contain a detailed comparison to Ghorbani et al (see reply to #1), acknowledging the intuition contained in the figure. (Sanity Checks for Saliency Maps) is already discussed in the text. Bounding the (local) Lipschitz constant of the explanation has the disadvantage that it makes the explanation insensitive to *any* small perturbation, e.g. even those which lead to a substantial change in output. This is clearly undesirable as the heatmap should explain why the perturbation leads to such a drastic "change of mind" of the network. Our method does not have this problem, since it only bounds the same output curvature. The final ms will explain the relation to the nice work by (Alvarez-Melis, Jaakkola) in detail. We fully agree with the indicated relation to PGD. The same 100 images were used for comparability. We implemented all your suggestions.

Reply to #4: Since our adv attacks move on lines of the same output, they are orthogonal to conventional adv attacks on the classification. They are therefore (locally) independent. We conducted experiments confirming this theoretical prediction (cosine of angle between perturbations averaged over 100 images is $-1.6 \times 10^{-9} \pm 1.6 \times 10^{-8}$). We added pixel flipping results which show that there is indeed a trade-off between robustifying explanations and their performance. For $\beta \ll 1$, the method is provably more robust but does not perform well. Our choice $\beta \approx 1$ however lies in a sweet spot leading to better explanations *and* robustness. We implemented all your suggestions.



(a) Heat on Truck (b) Heat on Street (c) Heat on Sky (d) Quantitative Analysis (e) Pixel Flipping

Figure 1: (a-c) Our method can structurally reproduce a heatmap, Ghorbani et al cannot. We use the same image as Ghorbani et al for comparability (d) Similarity to target map (higher is better) and change of winning-class probability for 100 images. (e) β-smoothing leads to superior pixel flipping performance (smaller is better).