Explaining Image Classifiers by Counterfactual Generation

Chun-Hao Chang, Elliot Creager,
Anna Goldenberg, David Duvenaud

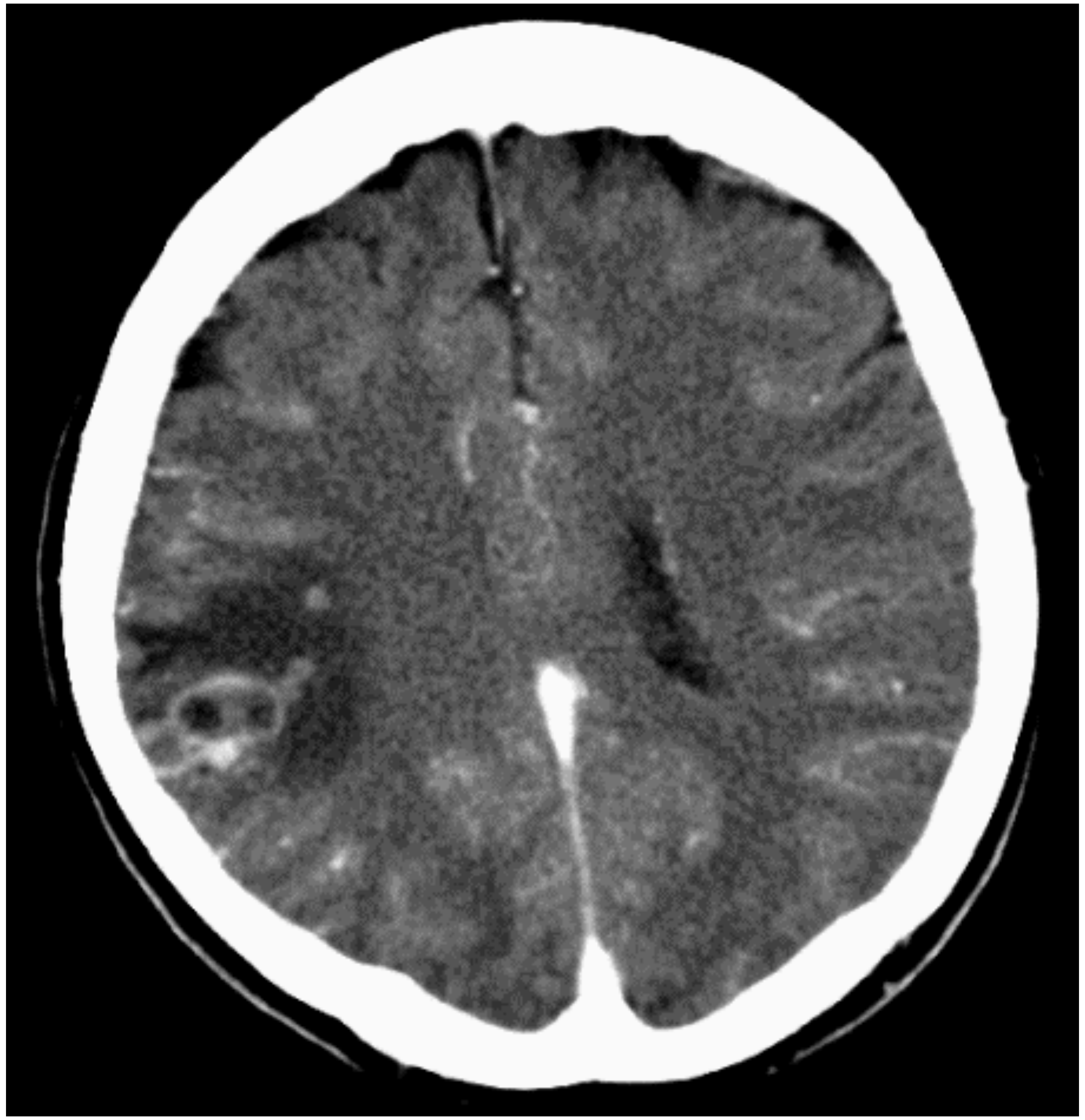
What is an explanation?

- Something that, if it had been different, would have changed the answer.

- “Why”: the cause, reason, or purpose for which

- “Why was X true?” -> “What, if it had been different, would have made X not true?”

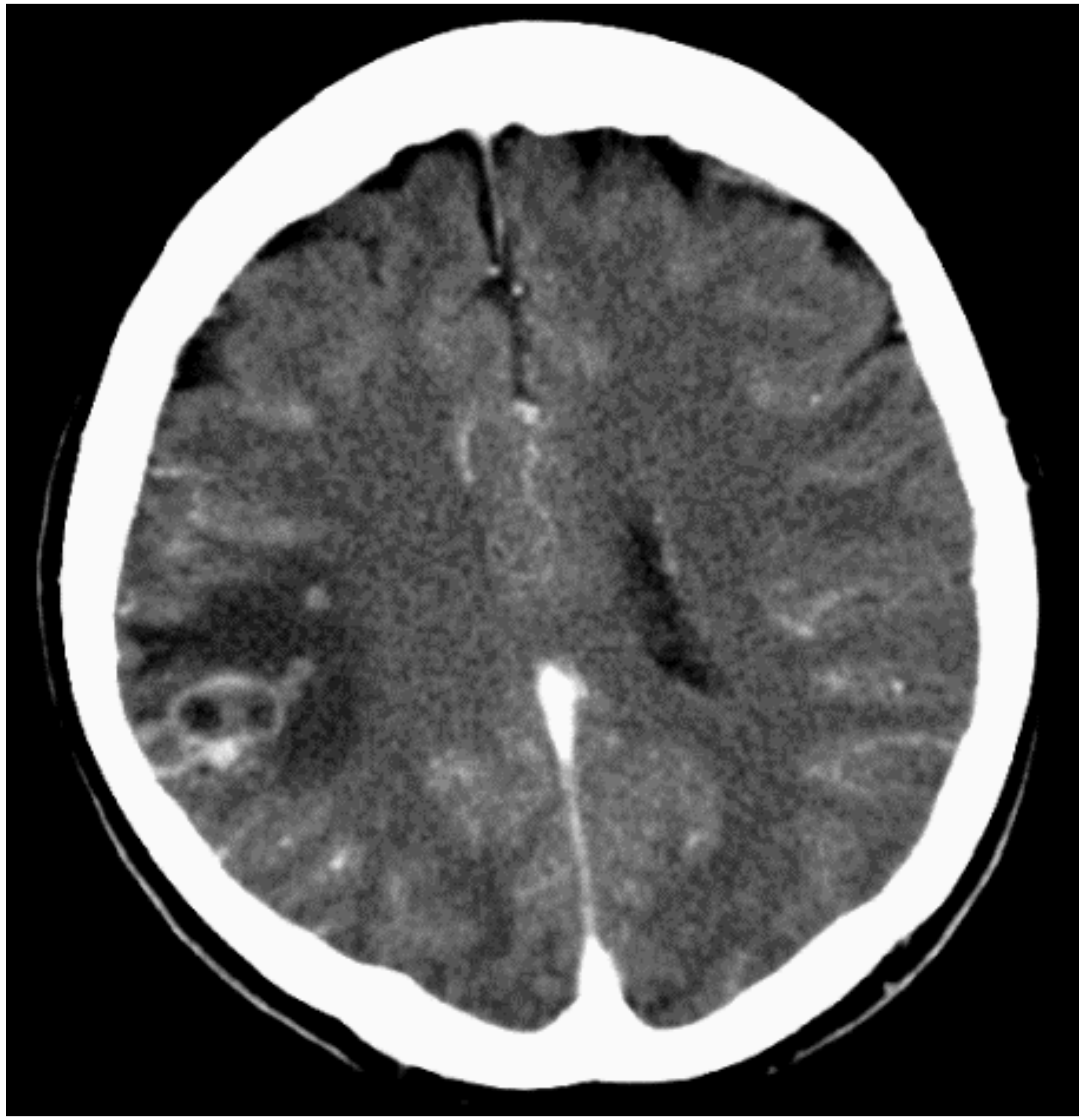
- Example: This part of the image makes me think it’s cancer. If it had been the usual color, I wouldn’t have a reason to worry.



What is an explanation?

- "Why was X true?" -> "What, if it had been different, would have made X not true?"

- Many possible answers, would like to prioritize plausible alternatives.



How to automate explanation?

- Need:

  1. Automatic answer-giver (i.e. image classifier) p(y|x)

  2. Automatic source of plausible counterfactuals p(x)

- Can ask: “What about this image, had it been different, would have changed the classification”

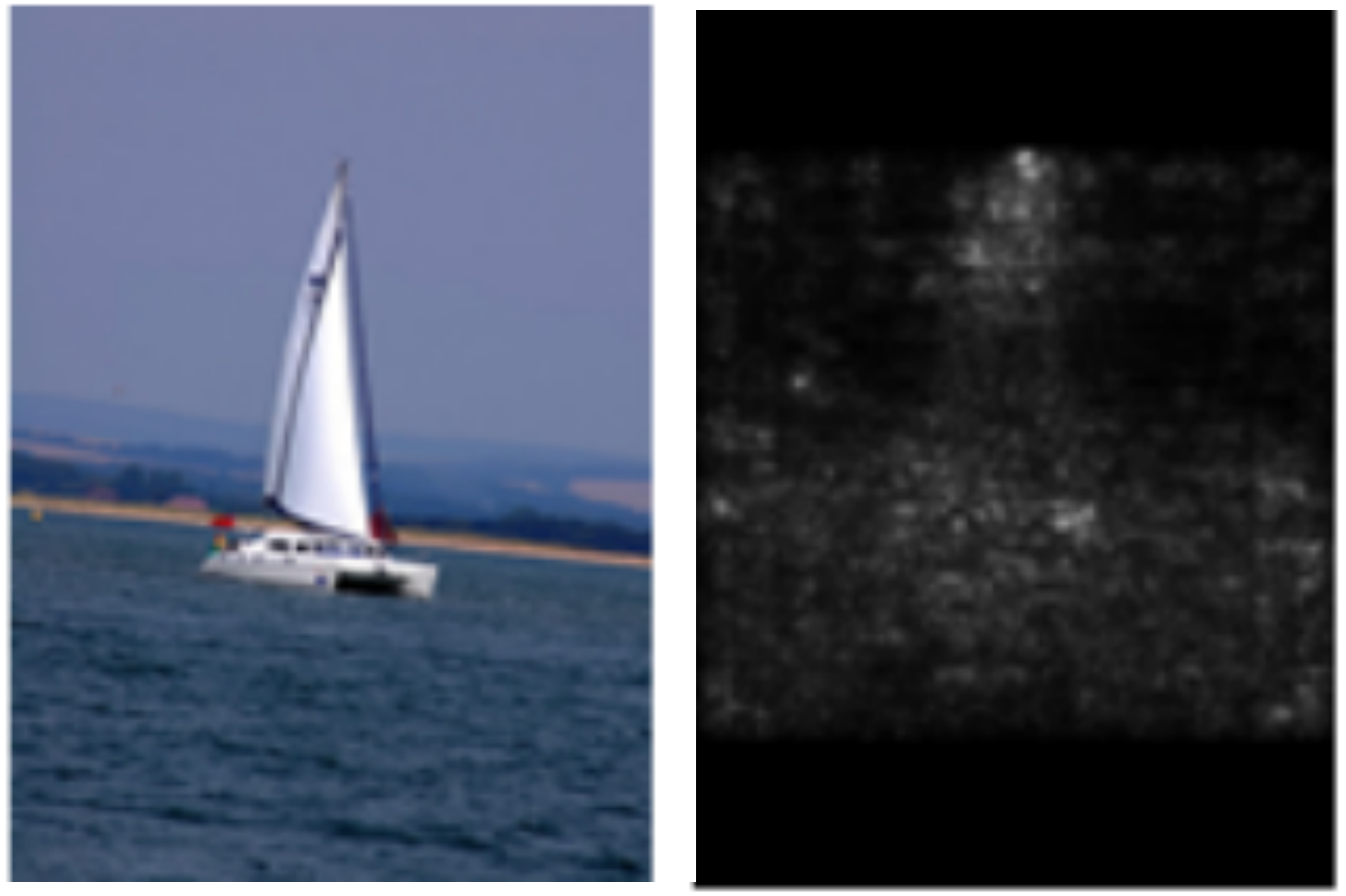
Previous work

- Original “saliency maps” simply plot gradient:

$$\frac{\partial}{\partial x} p(c \mid x)$$

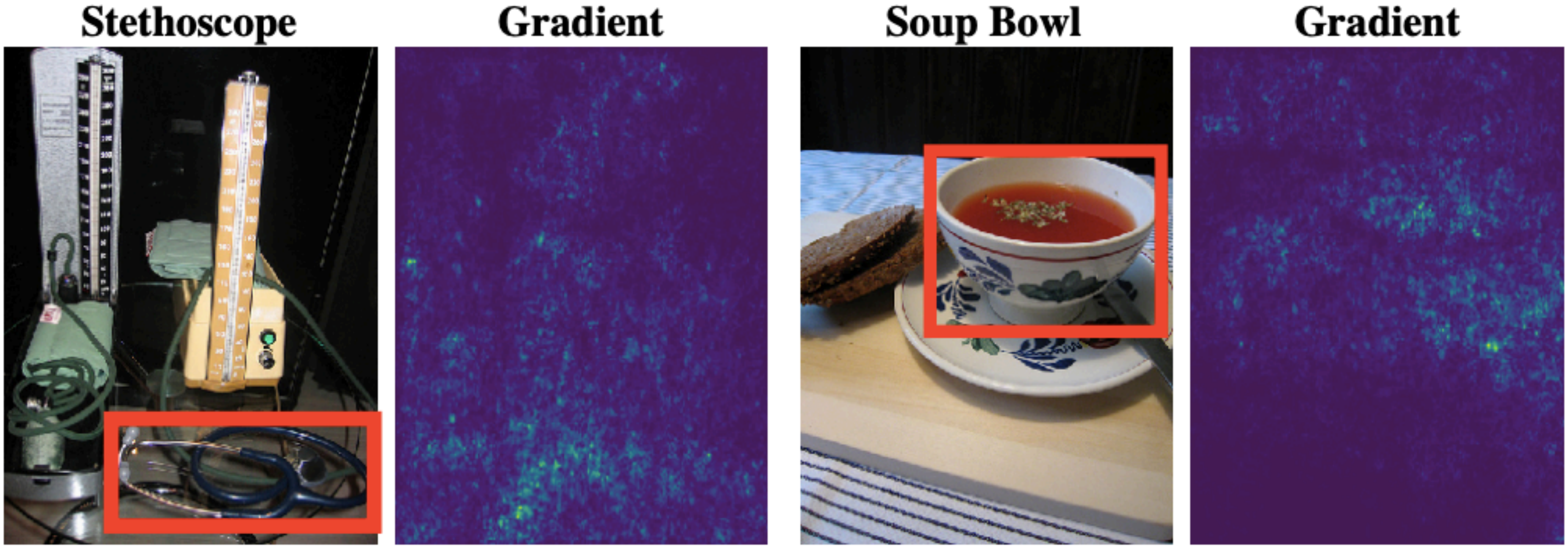
- Answers question: Which direction of change in pixels would most change the label?

- A sort of instantaneous counterfactual.



Simonyan et al., 2014

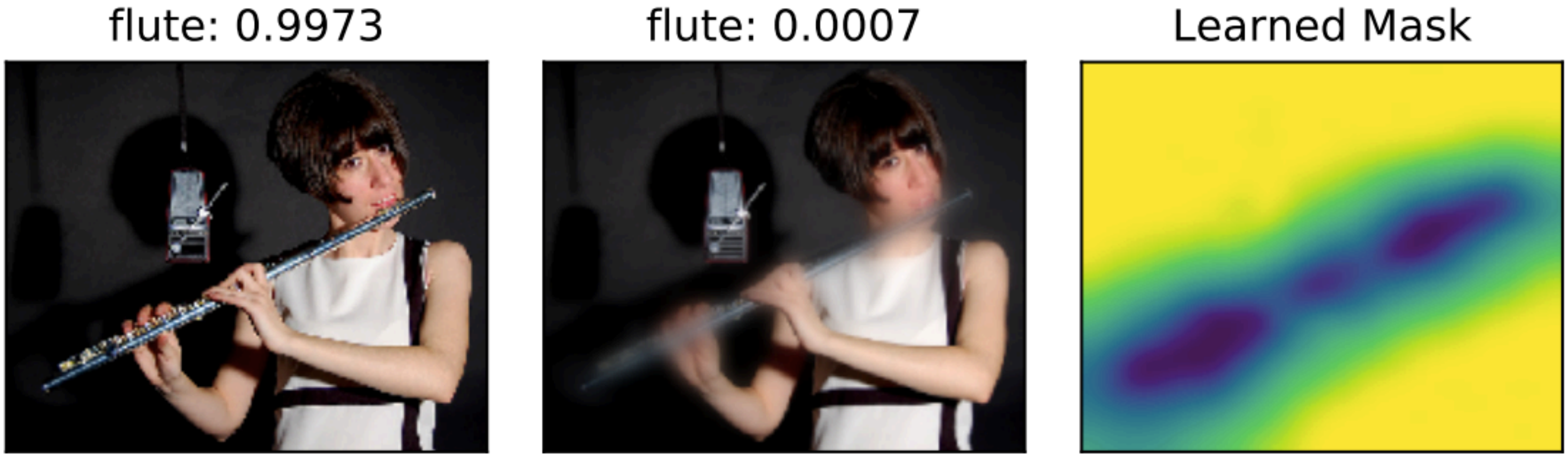
Saliency maps ask wrong question



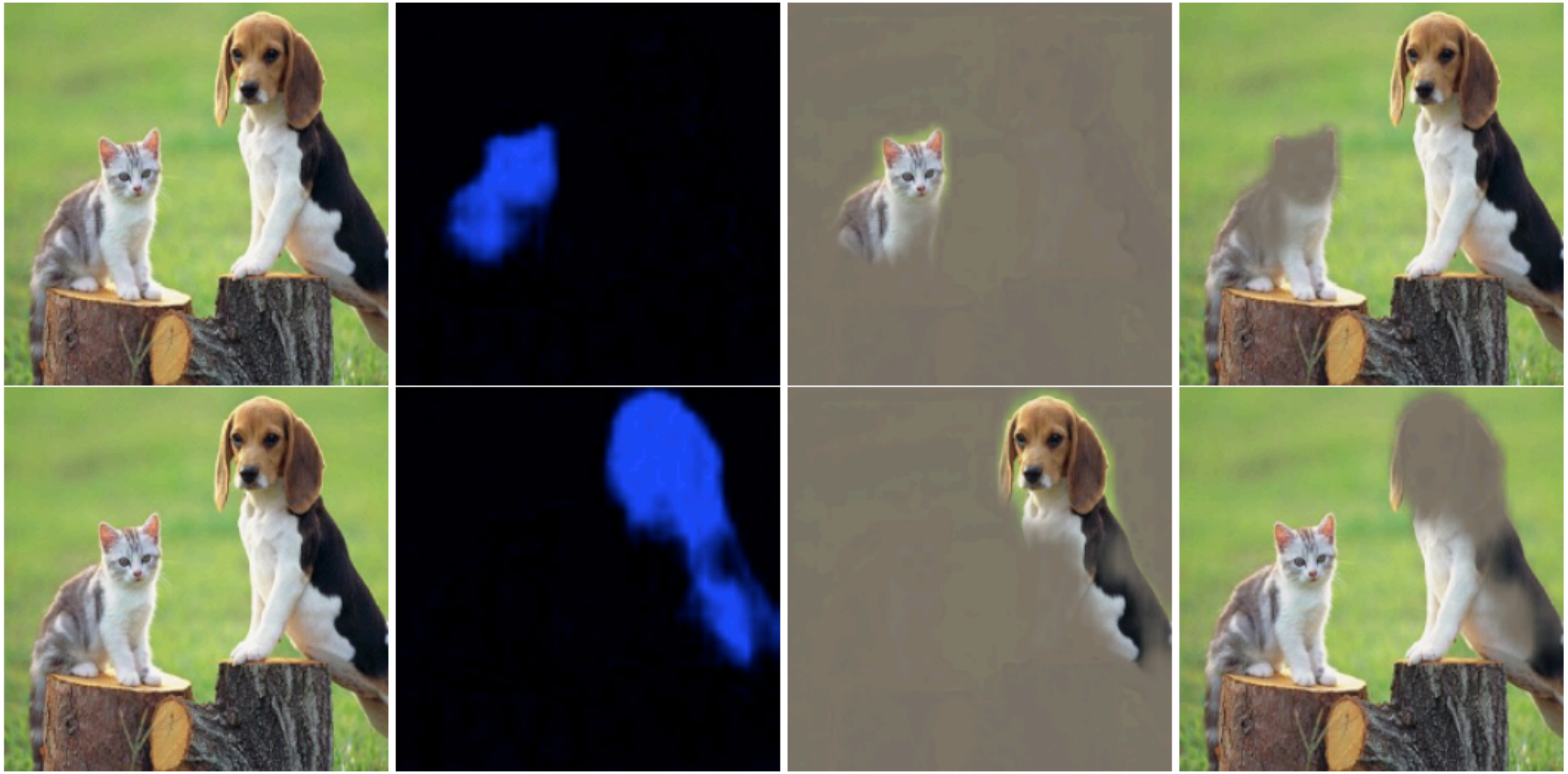
- Fong & Vedaldi, 2017

Related work

- Gradient maps have weird artefacts, related to adversarial examples.
- Fong & Vedaldi, 2017 ask which parts must be blurred



Dabkowski and Gal, 2017



(a) Input Image (b) Generated saliency map (c) Image multiplied by the mask (d) Image multiplied by inverted mask

Our approach

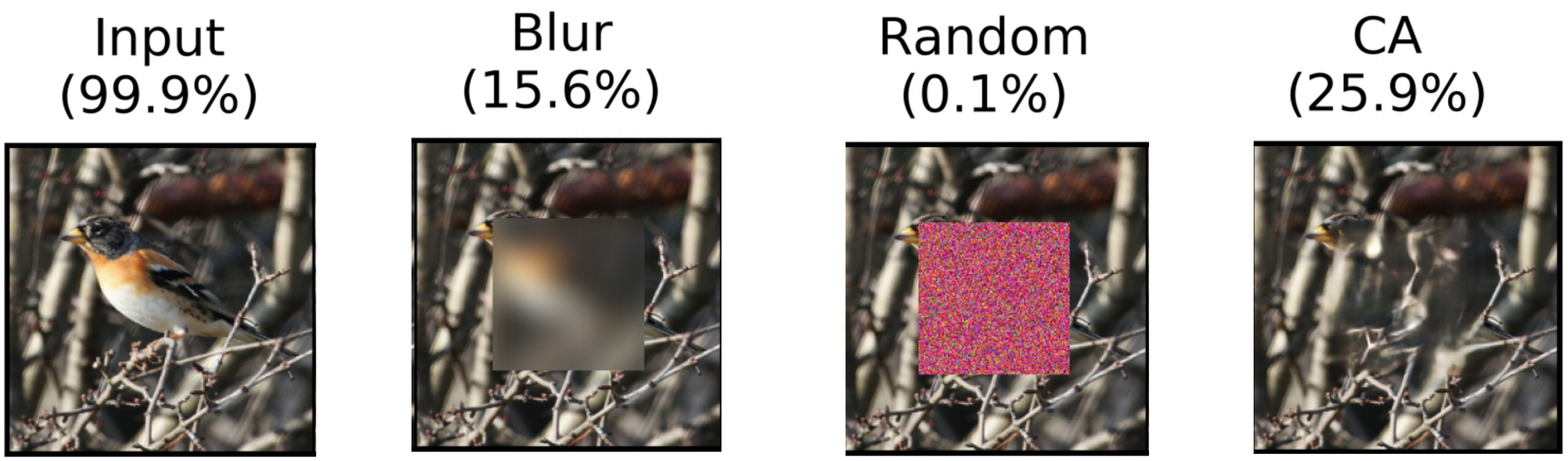
- Existing method's counterfactuals are based on implausible alternatives.

- We ask: "What region of this image, had it not been seen, would most have changed the classification"

- Fill in with consistent, plausible alternative image patches

$$p_{\mathcal{M}}(c|\boldsymbol{x}_{\backslash r}) = \mathbb{E}_{\boldsymbol{x}_r \sim p(\boldsymbol{x}_r|\boldsymbol{x}_{\backslash r})} \left[p_{\mathcal{M}}(c|\boldsymbol{x}_{\backslash r}, \boldsymbol{x}_r) \right]$$

Conditional Counterfactual Generation



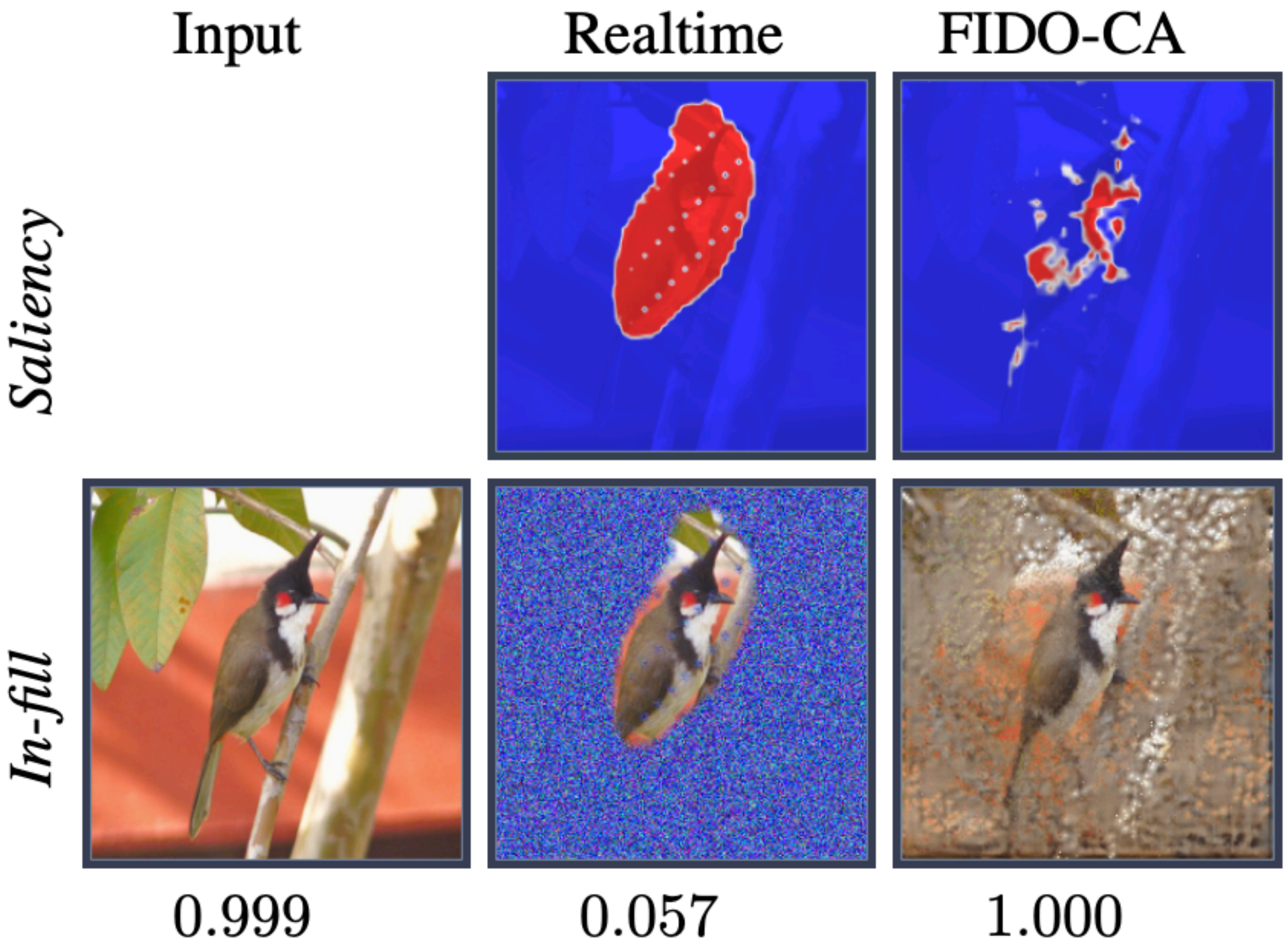
- For image classifiers, need to generate plausible alternative in-fills of images.
- Can use variational autoencoders, or GANs.
- Sum over all possible in-fills: $p_{\mathcal{M}}(c|\boldsymbol{x}_{\backslash r}) = \mathbb{E}_{\boldsymbol{x}_r \sim p(\boldsymbol{x}_r|\boldsymbol{x}_{\backslash r})}\left[p_{\mathcal{M}}(c|\boldsymbol{x}_{\backslash r}, \boldsymbol{x}_r)\right]$

The converse question

- Can also ask: “Which part of the image, if the rest were obscured, would keep the class the same?”

- I.e. what are non-essential parts of the image. Aka Smallest Deleted Region (Dabkowski and Gal, 2017)

- Our method (FIDO): Optimize to mask out as much of image as possible while keeping counterfactual answer same.



Details of approach

- Optimize soft mask

$$L_{SDR}(\theta) = \mathbb{E}_{q_\theta(\boldsymbol{z})}\left[s_{\mathcal{M}}(c|\phi(\boldsymbol{x}, \boldsymbol{z})) + \lambda\|\boldsymbol{z}\|_1\right]$$

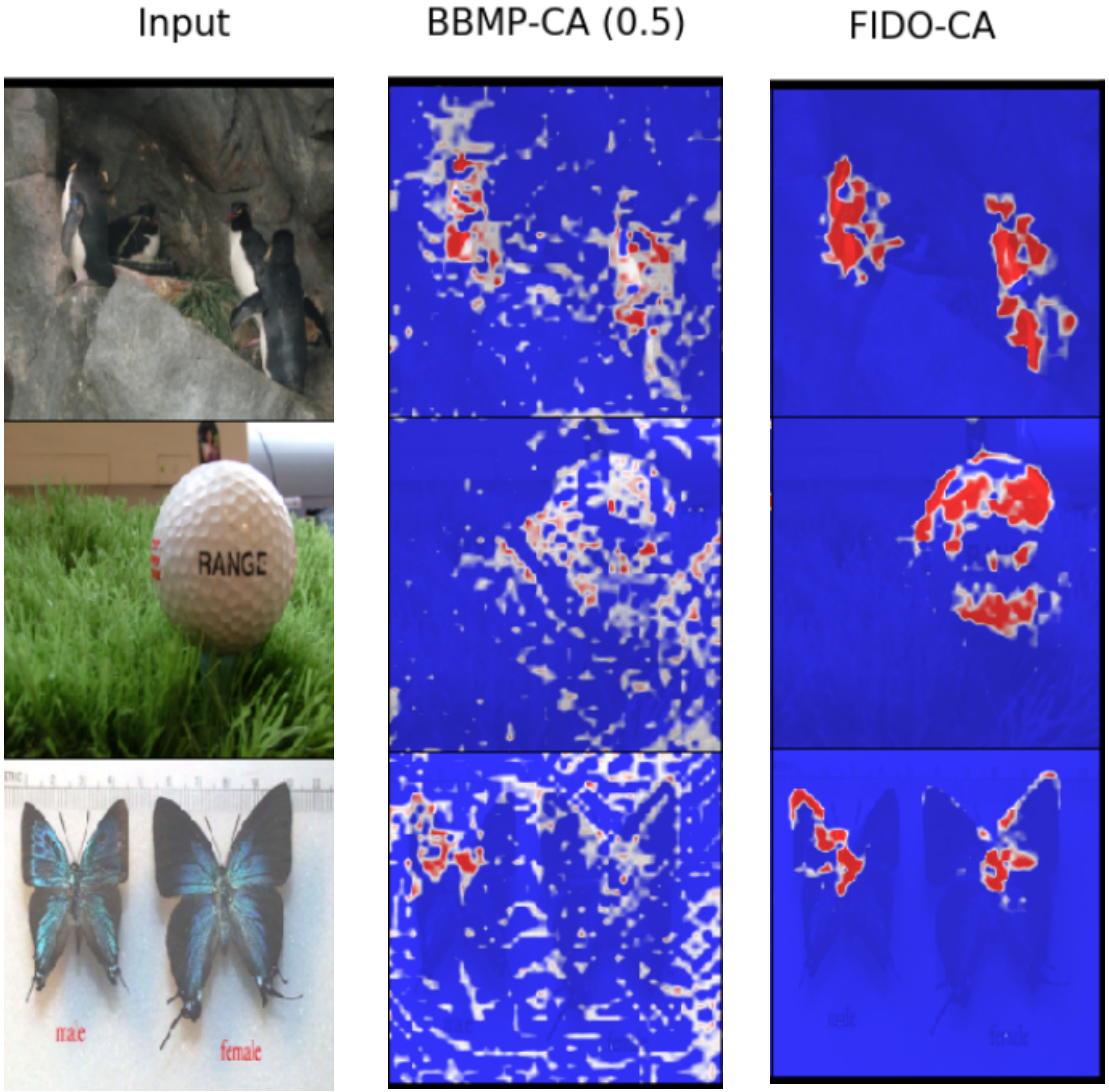
- Integrate over possible infills in inner loop with Monte Carlo

$$s_{\mathcal{M}}(c|\boldsymbol{x}) = \log p_{\mathcal{M}}(c|\boldsymbol{x}) - \log(1 - p_{\mathcal{M}}(c|\boldsymbol{x}))$$

- Require sparsity penalty

$$q_\theta(\boldsymbol{z}) = \prod_{u=1}^{U} q_{\theta_u}(z_u) = \prod_{u=1}^{U} \mathrm{Bern}(z_u|\theta_u).$$

Qualitative Results



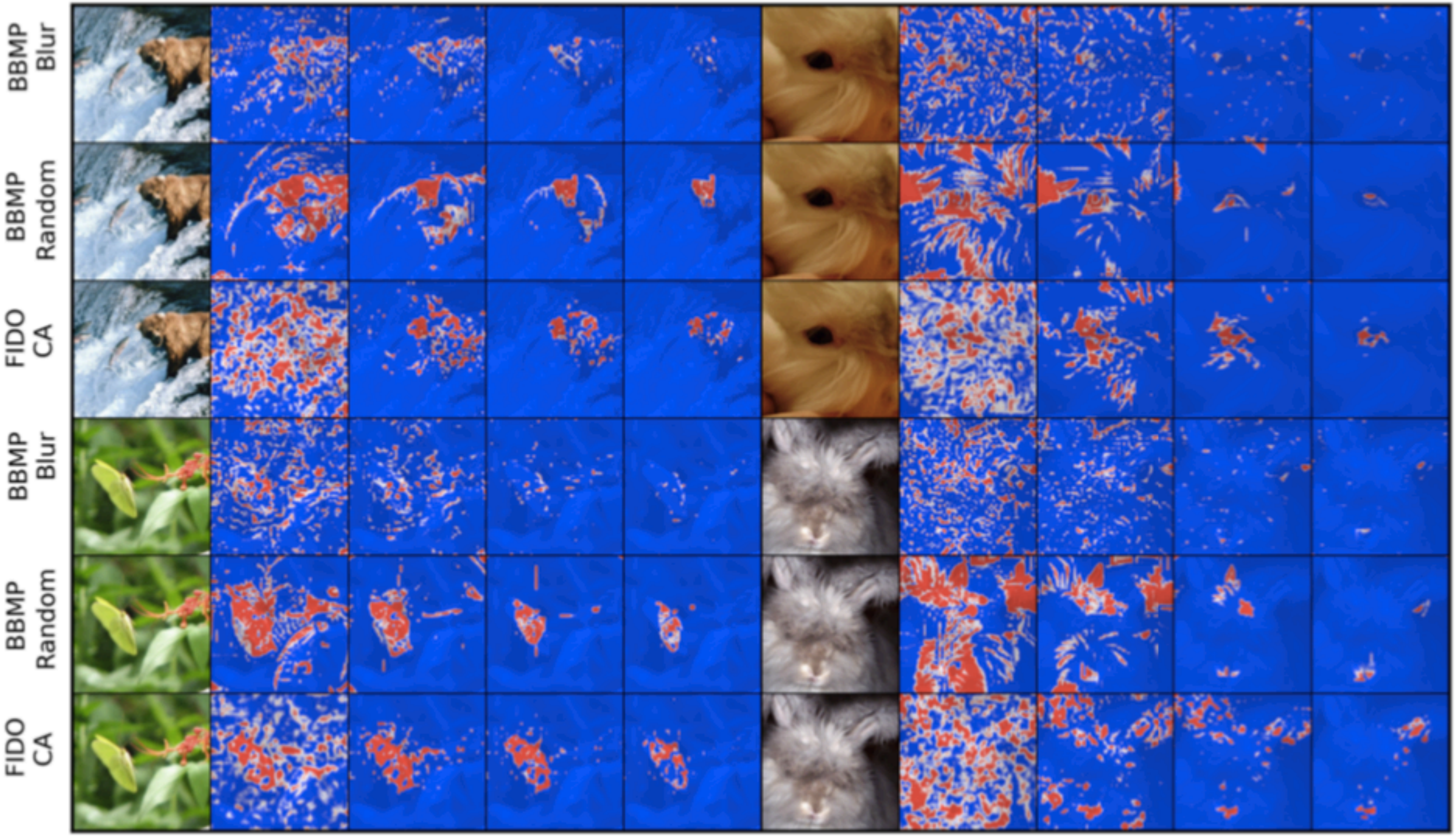


Figure 7: **BBMP vs FIDO saliency map** by increasing the sparsity penalty λ value from left to right.

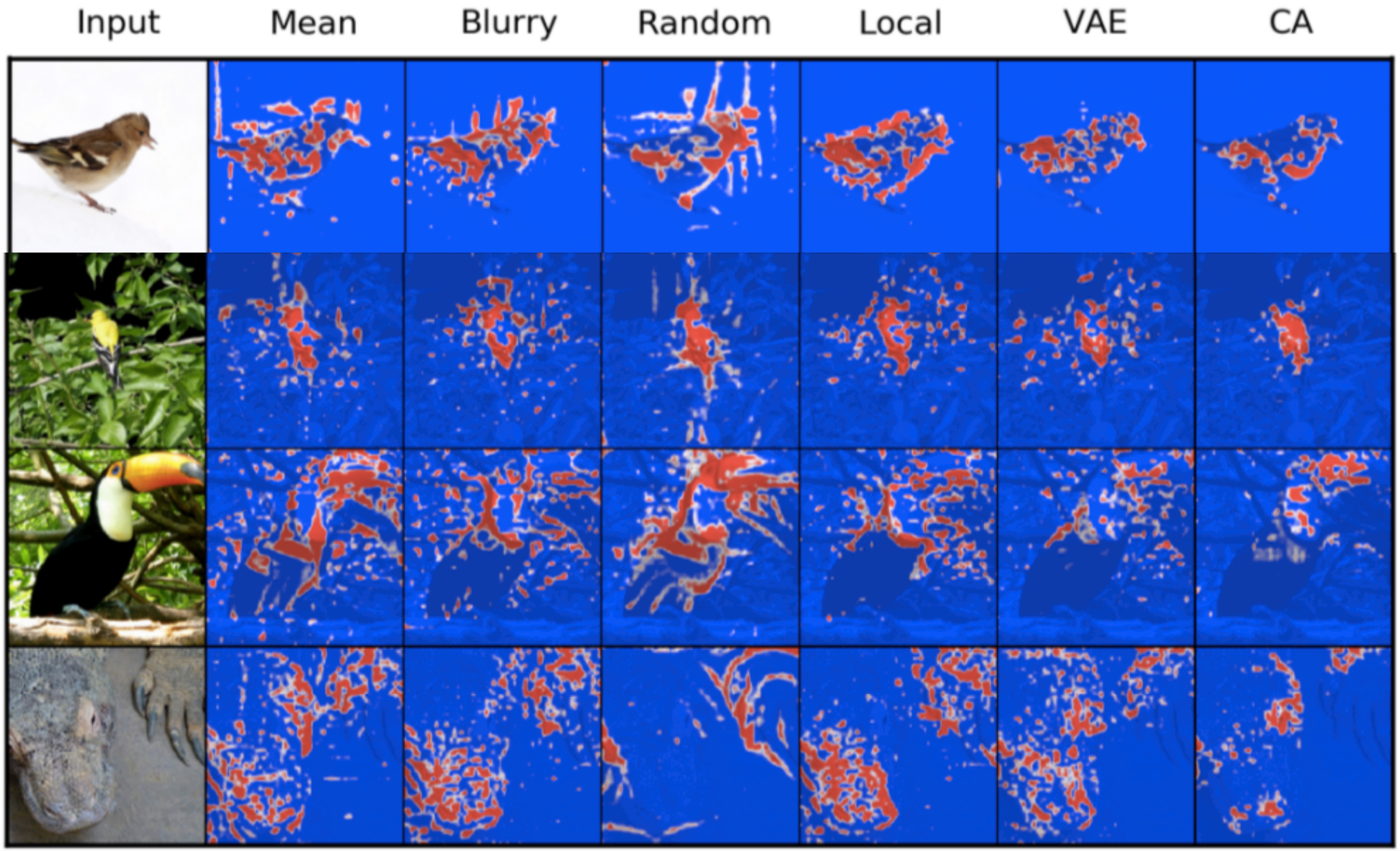
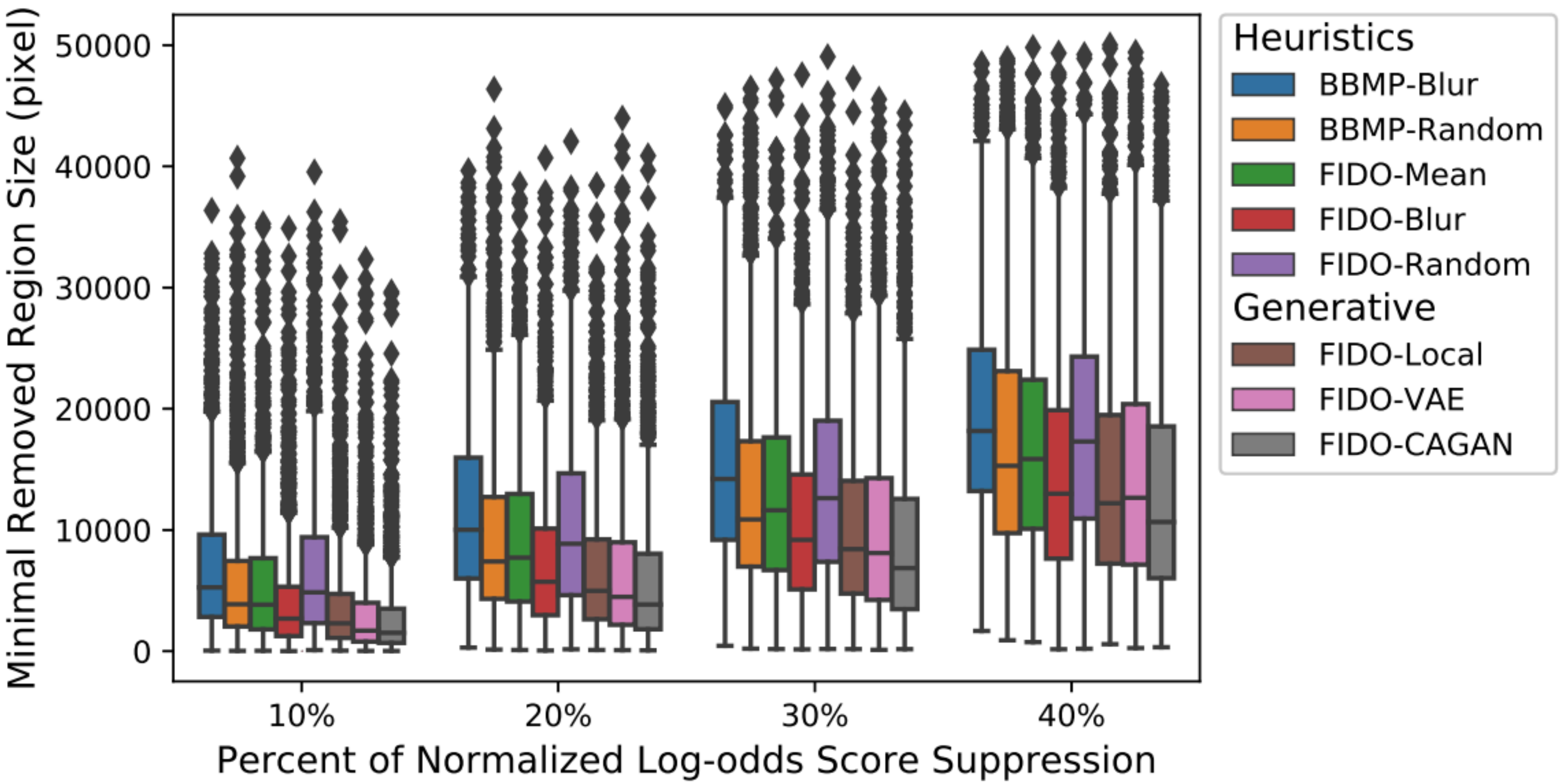


Figure 4: **Comparison of saliency map under different infilling methods** by SSR using ResNet.



Number of salient pixels required to change normalized classification score.

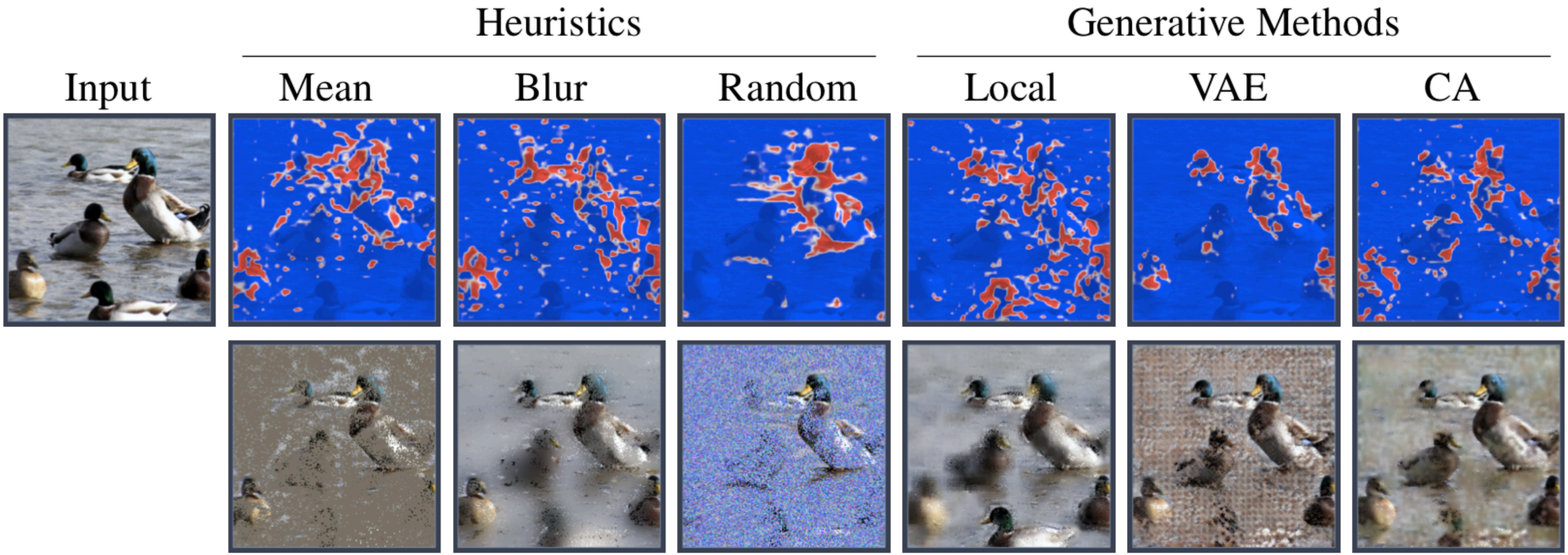
Technical Limitations

- Quality of conditional generative models. GANs are good at generation, still hard to condition on part of the image.

- Speed of approximate inference
(necessary for fast infilling)

- Optimization over shape of masked region. Would prefer hard mask edges.

Progress in Generative Models Needed



Conceptual Limitations

- Parts of images are a blunt tool for explanation. Better answers in terms of higher-level latent variables?
- Should probably offer multiple explanations
- Should probably relate explanations to actions that can be taken by the user.

Higher-level Counterfactuals

ablate person units

ablate curtain units

ablate window units

ablate table units

Bau et al 2019

Takeaways

- Conditional generative models let us automatically reason about counterfactuals
- Figuring out what question to ask is the hard part!

Thanks!