Saliency Strikes Back: How Filtering out High Frequencies Improves White-Box Explanations

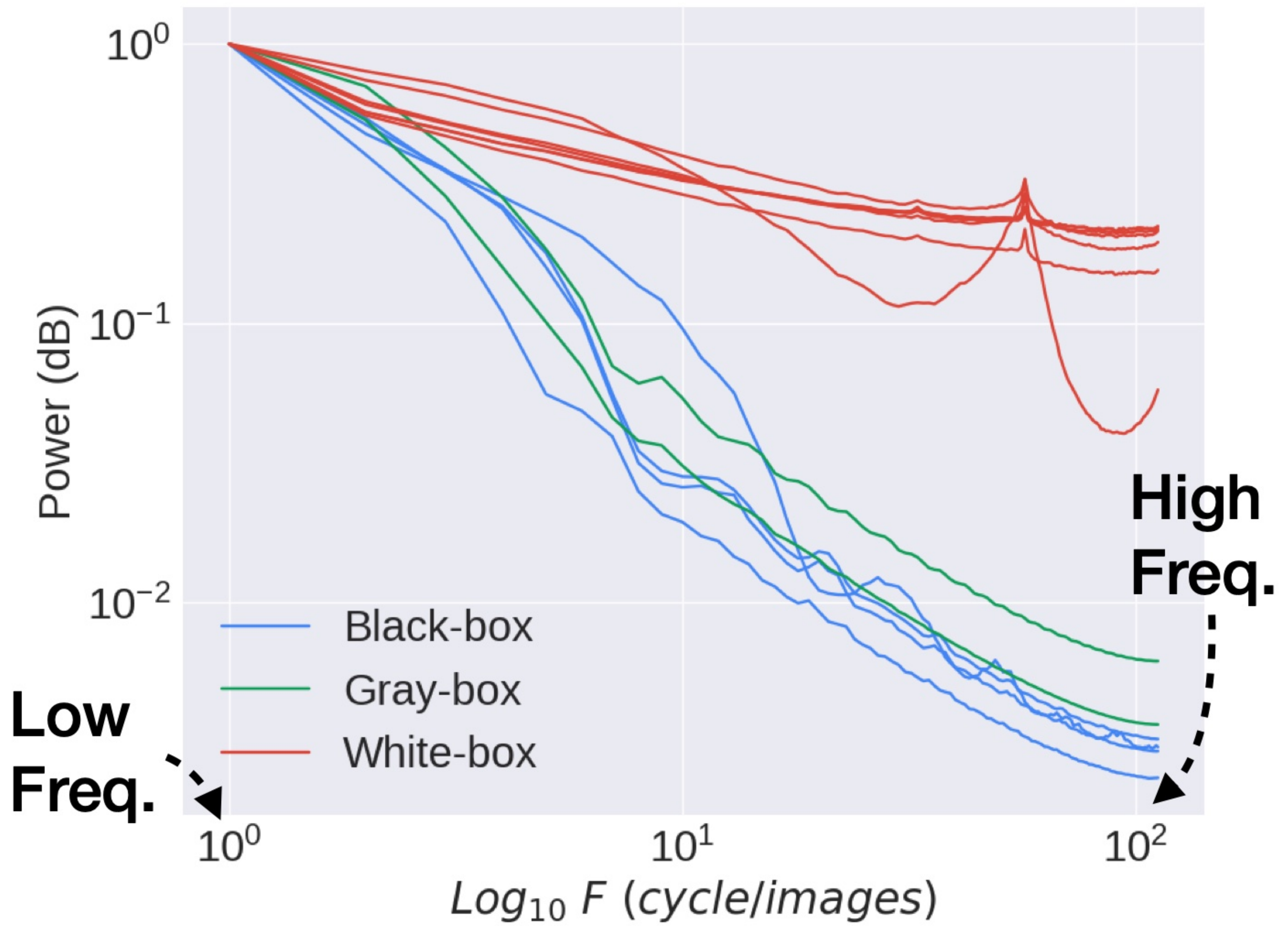
Sabine Muzellec[1,2], Thomas Fel[1,3], Victor Boutin[1,2], Léo Andéol[3,4], Rufin VanRullen[2], Thomas Serre[1]

[1]Carney Institute for Brain Science, Brown University, USA [2]CerCo, CNRS, France
[3]SNCF, France [4]Institute of Mathematics of Toulouse, University Paul Sabatier, France

ICML 2024 - July 21st to 27th - Vienna, Austria

Comparing attribution methods

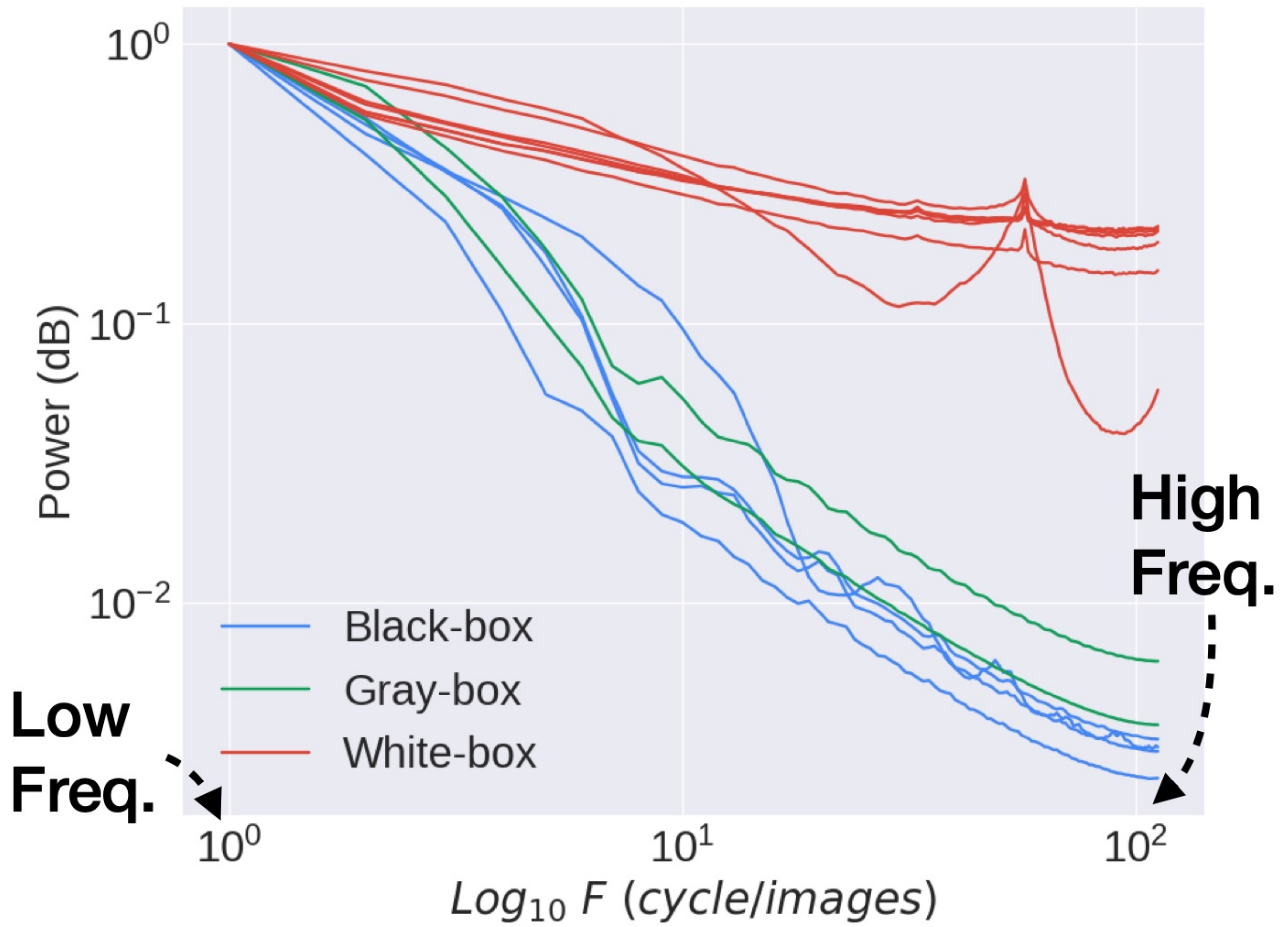
Using their Fourier signature



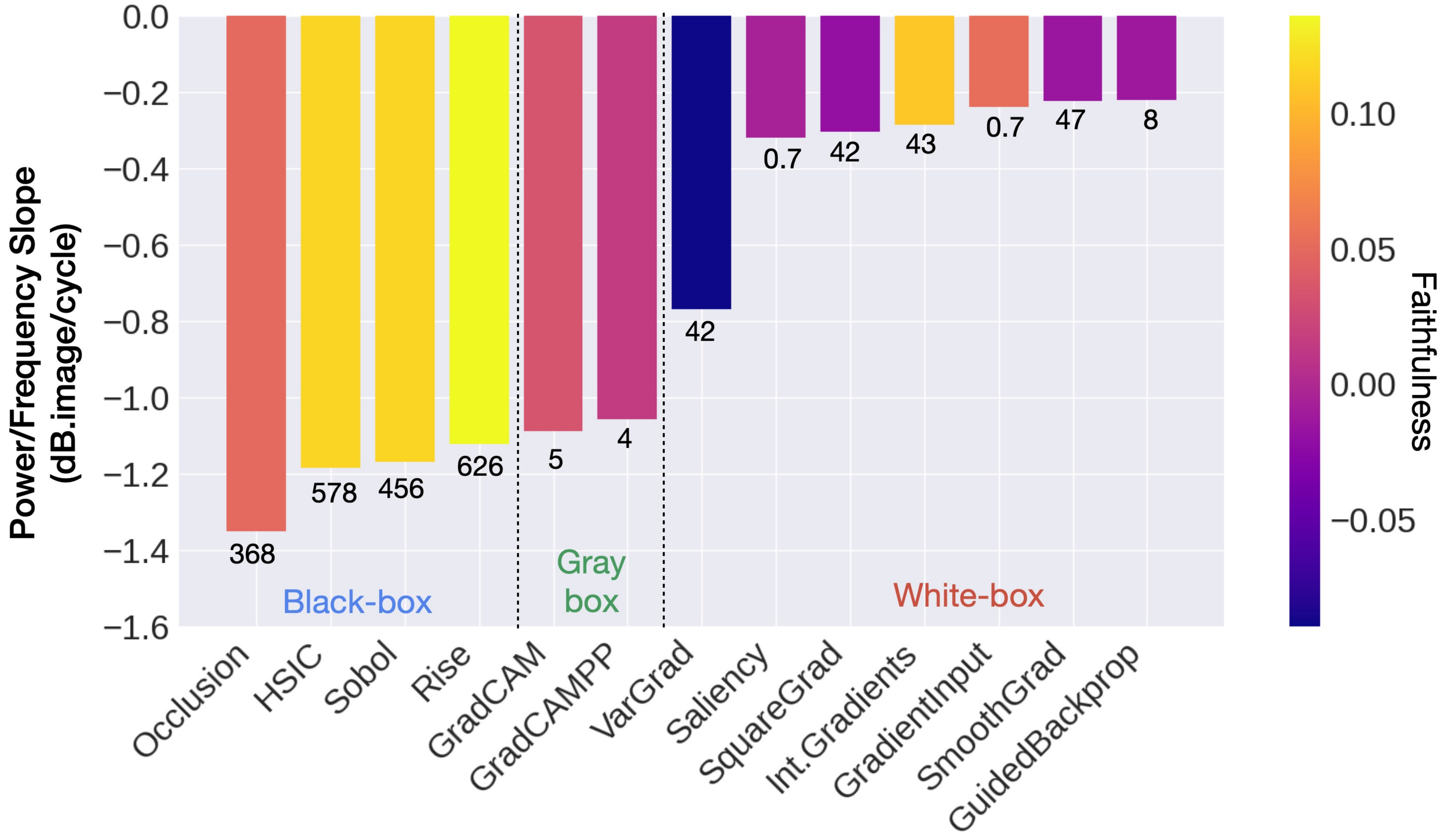
White-box attribution methods produce attribution maps with increased power in the high frequencies

Comparing attribution methods

Using their Fourier signature



White-box attribution methods produce attribution maps with increased power in the high frequencies



White-box methods are computationally more efficient but have lower faithfulness

What are these high-frequencies and where do they come from?

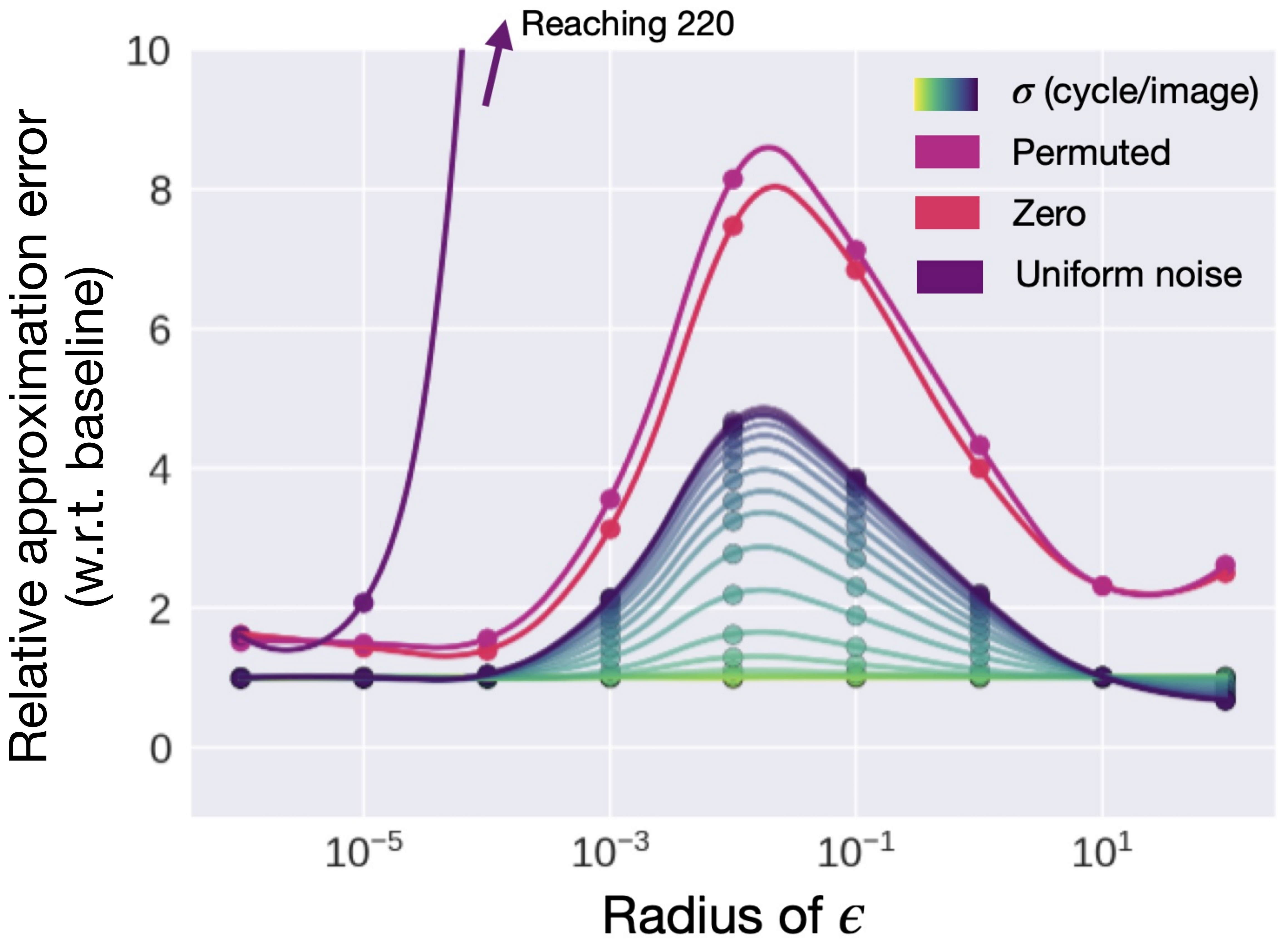
High-frequencies are artifacts

Considering:

$$f(x+\epsilon) \approx f(x) + \epsilon \nabla_x f(x) \qquad (1)$$

Given a cutoff frequency σ, we characterize the error between the Taylor expansion and the function through, $\zeta(x,\sigma)/\zeta(x,\sigma_{\text{max}})$, with $\sigma_{\text{max}} = 224$ (no filtering), through:

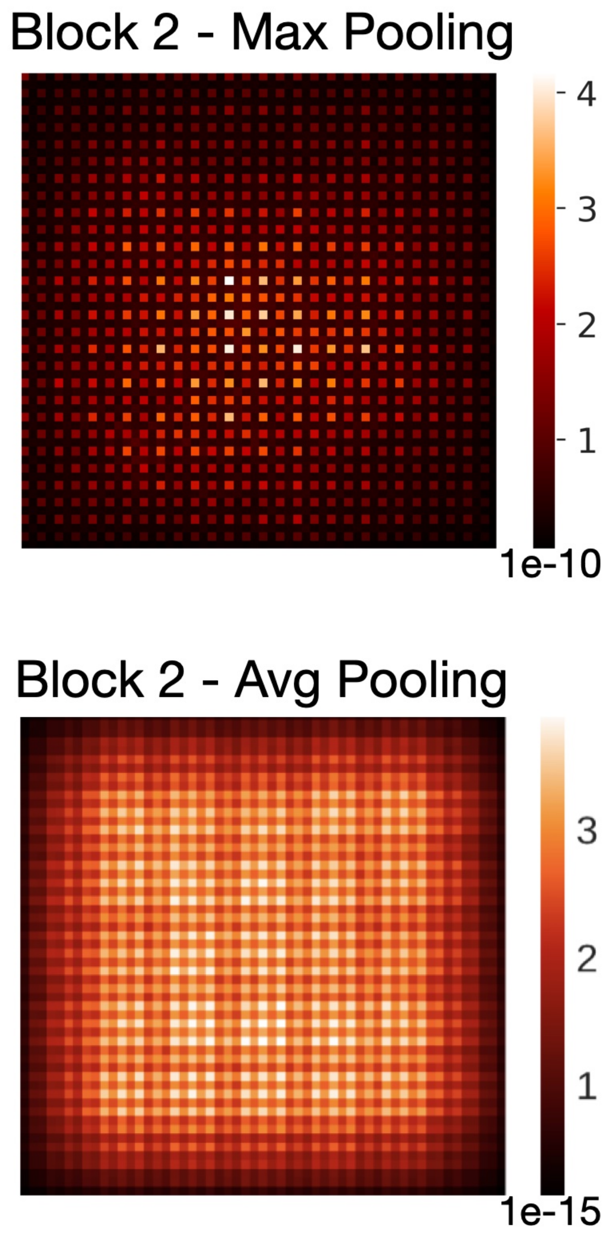
$$\zeta(x,\sigma) = ||f(x+\epsilon) - (f(x) + \epsilon \nabla_\sigma f(x))||_2$$

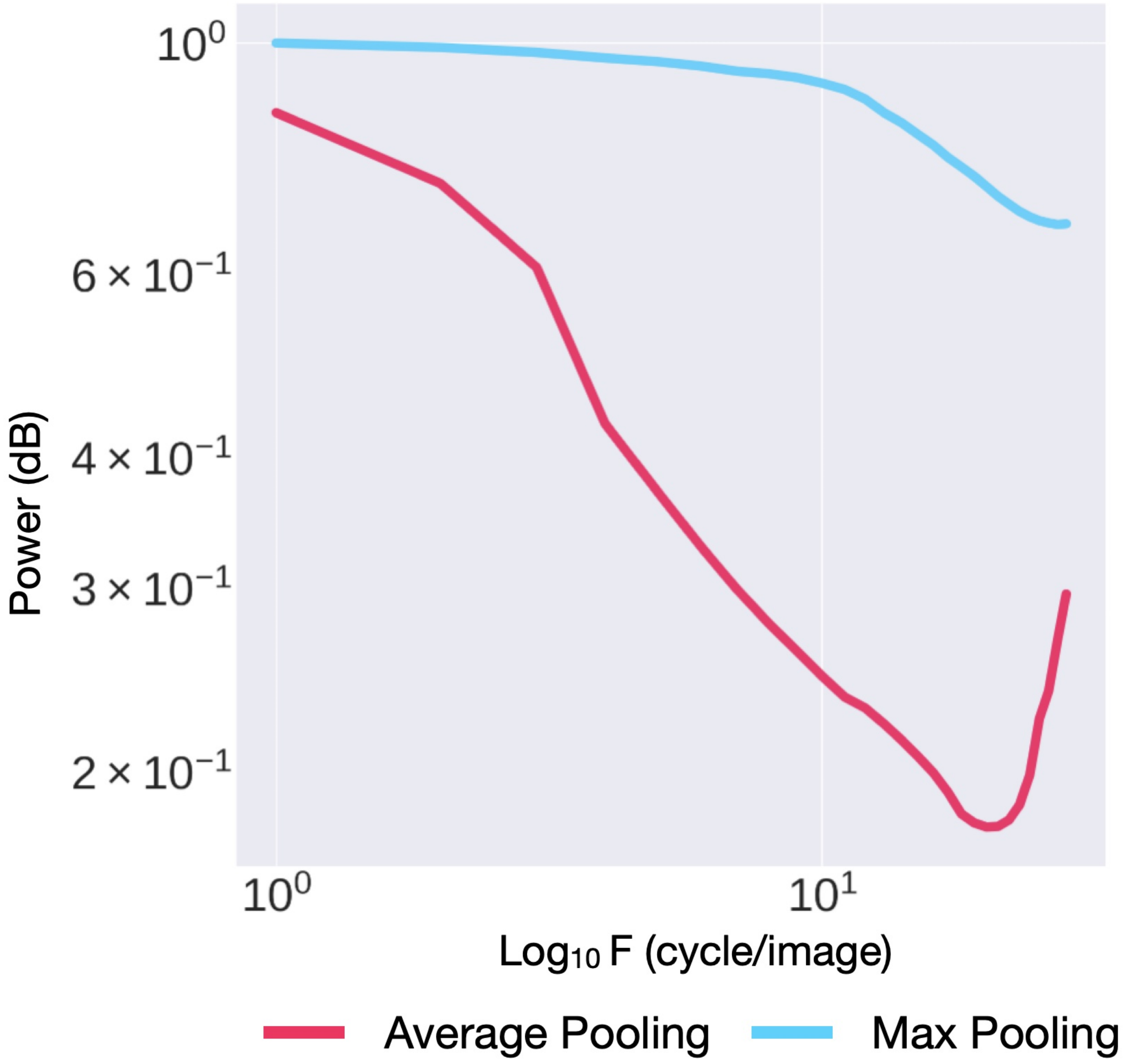


The filtered gradient still approximates the non-filtered gradient well when defined as the first-order term of a Taylor expansion.

And stem from Max Pooling operations

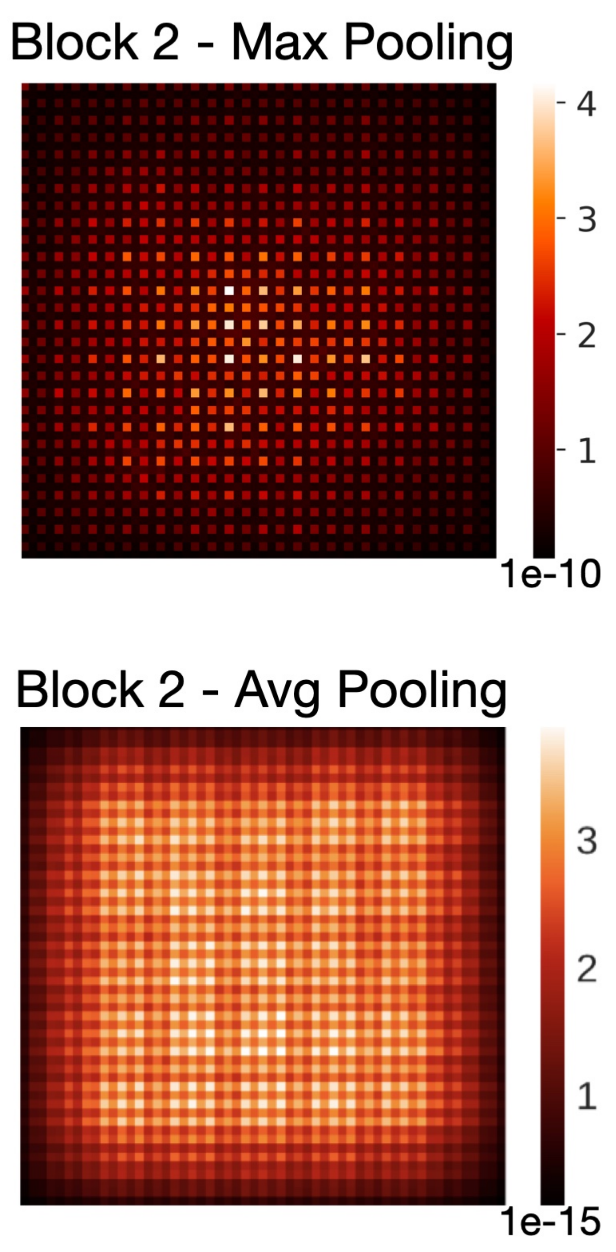
- The gradients following a Max Pooling operation exhibit checkerboard patterns.
- The Fourier signatures of the gradients resulting from a Max Pooling show more power in the high frequencies than those resulting from an Average Pooling.

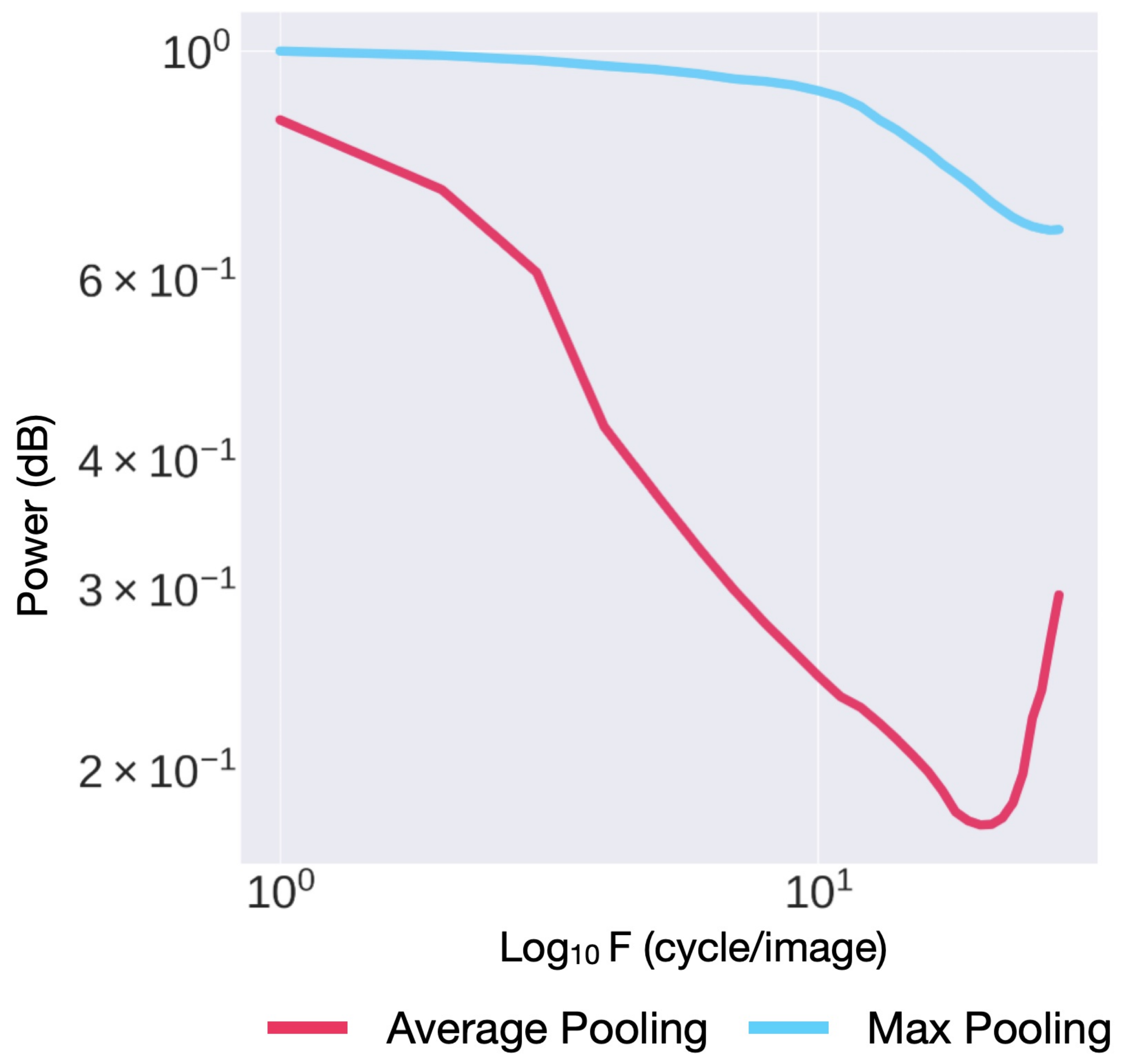


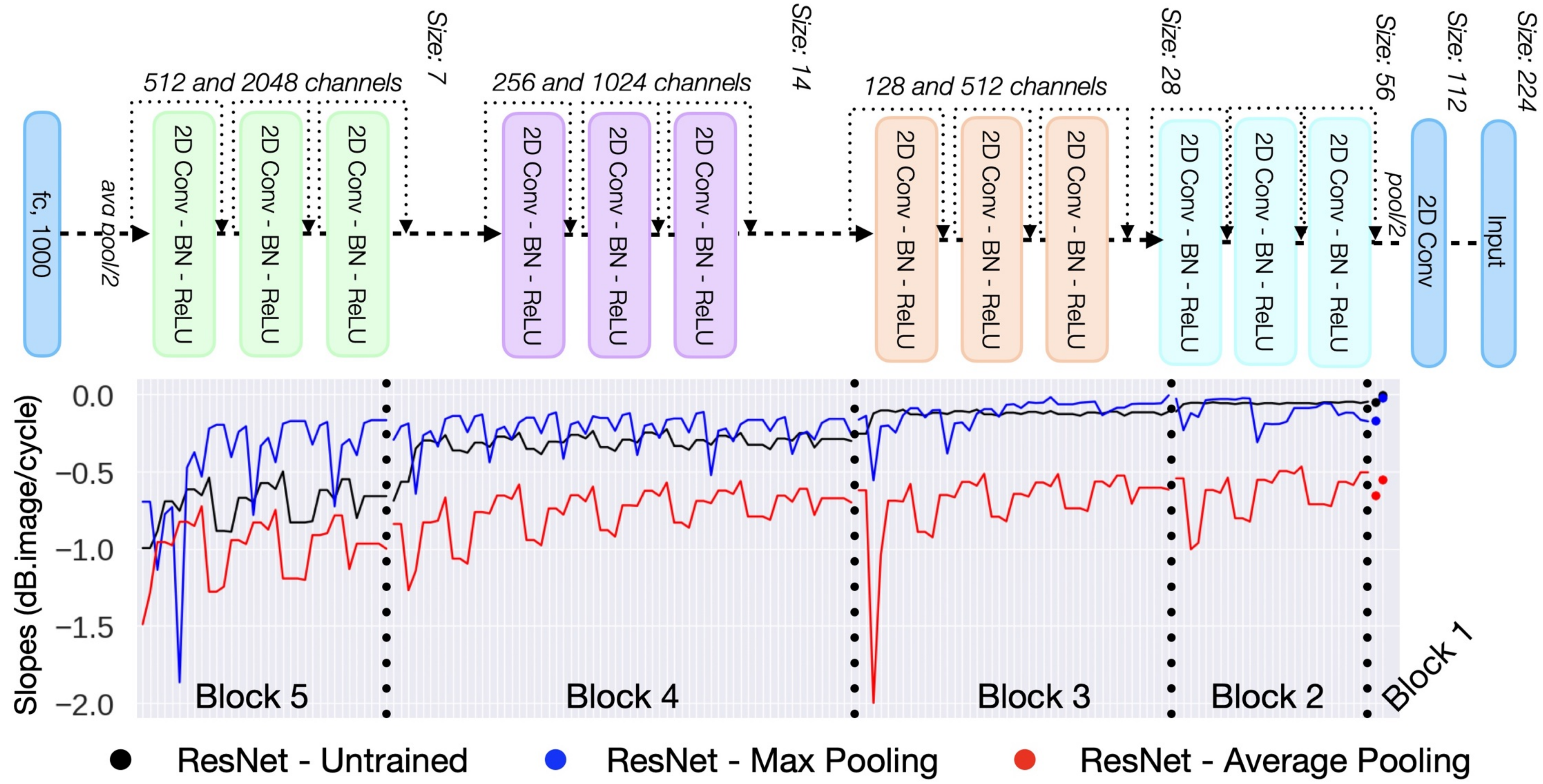


And stem from Max Pooling operations

- The gradients following a Max Pooling operation exhibit checkerboard patterns.
- The Fourier signatures of the gradients resulting from a Max Pooling show more power in the high frequencies than those resulting from an Average Pooling.
- This effect is cumulative over the depth of the model, and is not alleviated by training.







Can we repair the white-box methods by low-pass filtering these artifacts?

FORGrad: Fourier Reparation of the Gradient

FORGrad:

- Selects the cutoff frequency, $\sigma^{\star}$, to maximize faithfulness

FORGrad: Fourier Reparation of the Gradient

FORGrad:

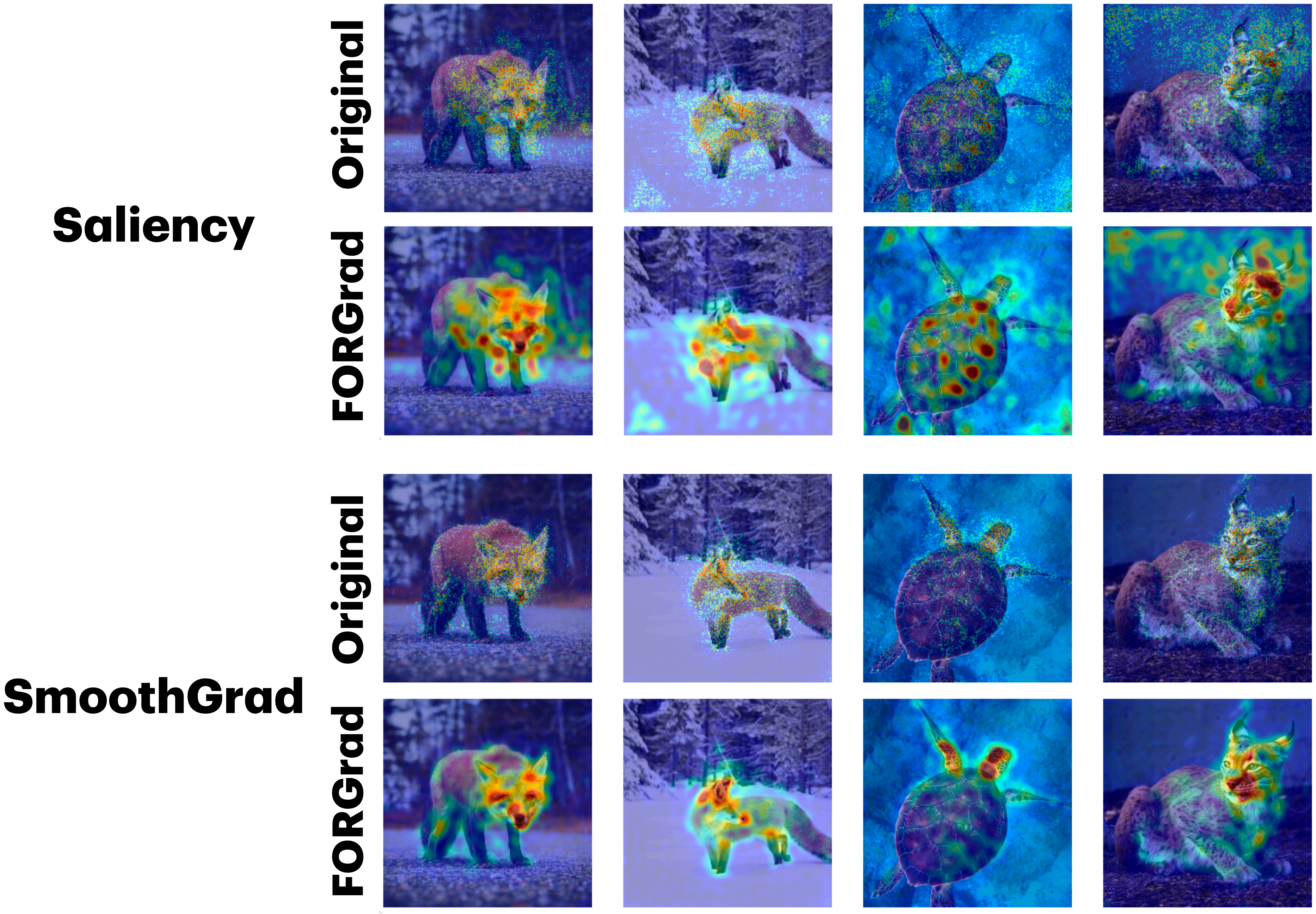
- Selects the cutoff frequency, $\sigma^\star$, maximizing the faithfulness
- Improves the score on all other metrics
- Can be applied to any architecture and attribution method

		ResNet50			
		Faith.(↑)	μFid.(↑)	Stab.(↓)	Time(↓)
White-box	Saliency(Simonyan et al., 2013)	0.18	0.40	0.67	0.78
	Saliency⋆	0.28	0.39	0.53	0.89
	Guidedbackprop(Ancona et al., 2018)	0.31	0.45	0.28	8.25
	Guidedbackprop⋆	0.35	0.45	0.22	7.05
	GradInput(Shrikumar et al., 2017)	0.2	0.36	0.42	**0.73**
	GradInput⋆	0.26	0.36	0.35	<u>0.77</u>
	Int.Grad(Sundararajan et al., 2017)	0.24	0.39	0.72	42.7
	Int.Grad⋆	0.31	0.38	0.76	41.3
	SmoothGrad(Smilkov et al., 2017)	0.23	0.45	0.22	46.6
	SmoothGrad⋆	<u>0.37</u>	0.44	0.21	48.3
	VarGrad (Adebayo et al., 2018)	0.36	<u>0.46</u>	**0.003**	41.5
	VarGrad⋆	0.35	0.44	<u>0.004</u>	40.6
	SquareGrad(Seo et al., 2018)	0.36	0.45	**0.003**	42.1
	SquareGrad⋆	0.36	<u>0.46</u>	0.005	40.9
Black & Gray-box	GradCAM(Selvaraju et al., 2017a)	0.31	0.40	0.31	5.24
	GradCAM++(Chattopadhay et al., 2018)	0.33	0.43	0.34	4.61
	Occlusion(Ancona et al., 2018)	0.20	0.39	0.6	368
	HSIC(Novello et al., 2022)	0.33	**0.47**	0.45	456
	Sobol(Fel et al., 2021b)	0.34	**0.47**	0.47	578
	RISE(Petsiuk et al., 2018)	**0.41**	0.34	0.55	626

FORGrad: Fourier Reparation of the Gradient

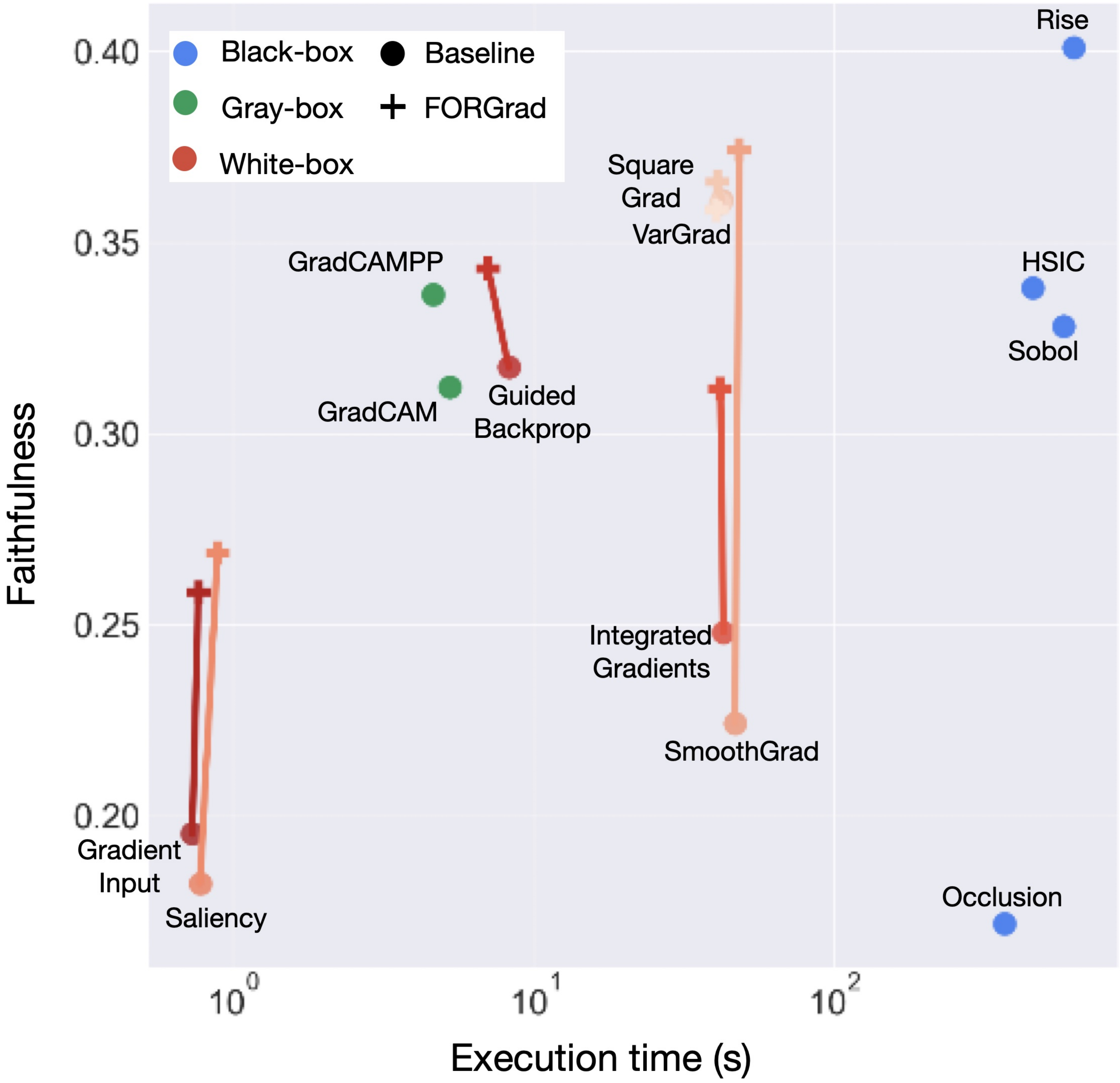
FORGrad:

- Selects the cutoff frequency, $\sigma^\star$, maximizing the faithfulness
- Improves the score on all other metrics
- Can be applied to any architecture and attribution method



Conclusion

- A major source of high-frequency artifacts in attribution maps computed with white-box methods is inherited from the model's gradients.
- These artifacts are a consequence of the max-pooling and striding operations used in convolutional neural networks (CNNs) and are responsible for the lower explainability scores of these methods.
- **FORGrad** filters out frequencies above a certain ideal cut-off value and systematically improves the explainability score of white-box methods while being significantly more computationally efficient.



Thank you!