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** [4 3] **Rufin VanRullen** [2] **Thomas Serre** [1]

Abstract

Attribution methods correspond to a class of explainability methods (XAI) that aim to assess how individual inputs contribute to a model's decision-making process. We have identified a significant limitation in one type of attribution methods, known as "white-box" methods. Although highly efficient, as we will show, these methods rely on a gradient signal that is often contaminated by high-frequency artifacts. To overcome this limitation, we introduce a new approach called "**FORGrad**." This simple method effectively filters out these high-frequency artifacts using optimal cut-off frequencies tailored to the unique characteristics of each model architecture. Our findings show that **FORGrad** *consistently enhances* the performance of existing white-box methods, enabling them to compete effectively with more accurate yet computationally more demanding "black-box" methods. We anticipate that, because of its effectiveness, the proposed method will foster the broader adoption of straightforward and efficient white-box methods for explainability, providing a better balance between faithfulness and computational efficiency.

1. Introduction

Understanding how modern Artificial Neural Networks (ANNs) make decisions remains a major challenge. Given the ever-increasing range of machine learning applications, the need for robust and reliable explainability methods is pressing (Doshi-Velez & Kim, 2017a; Jacovi et al., 2021). To meet this demand, eXplainable Artificial Intelligence (XAI) focuses on developing new tools to help users better understand how ANNs arrive at their decisions (Jacovi et al., 2021; Doshi-Velez & Kim, 2017b; Rudin, 2019). Among the wide range of XAI approaches, attribution methods have become the go-to approaches. They aim to identify the features that drive a network's decisions (Simonyan et al., 2013) by assigning them an importance score according to their contribution to the overall prediction. Attribution methods fall broadly into two main classes: "white-box" and "black-box" methods. White-box methods require full access to a network's information (weights, architecture, etc.) to understand which regions in the input are the most important for the model's decision. Typically, such methods rely on the gradient of the model's prediction w.r.t. the input to produce attribution maps. For instance, "Saliency", one of the earliest white-box methods, traces a model's decision back to its input image using such the gradient information (Simonyan et al., 2013). Although simple and computationally efficient, Saliency is also known to produce noisy and hard-to-interpret attribution maps (Nguyen et al., 2021; Kim et al., 2022; Colin et al., 2021; Serrurier et al., 2022; Hase & Bansal, 2020). Several extensions have been proposed to try to overcome these limitations, e.g., by varying slightly the input and accumulating the resulting gradient to obtain a smoother map (Simonyan et al., 2013; Zeiler & Fergus, 2014; Springenberg et al., 2014; Smilkov et al., 2017; Sundararajan et al., 2017; Zhou et al., 2016; Selvaraju et al., 2017a; Fong & Vedaldi, 2017).

Conversely, black-box methods focus on a model's predictions without explicitly requiring any gradient computation. Such methods aim to quantify how changes in the inputs affect a model's output by altering the input and evaluating the newly obtained prediction (Zeiler & Fergus, 2014; Petsiuk et al., 2018; Fel et al., 2021a; Ribeiro et al., 2016; Novello et al., 2022). These black-box methods are computationally much more demanding as they require hundreds of network queries (or forward passes), but they outperform white-box methods on current XAI benchmarks. We refer the reader to Section 2 for more details on the different types of attribution methods. Despite extensive research (Sixt et al., 2020; Adebayo et al., 2018; Slack et al., 2021; Han et al., 2022; Agarwal et al., 2022; 2021; Pukdee et al., 2023), a question remains: Why do black-box methods outperform white-box methods?

[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. Correspondence to: Sabine Muzellec <sabine_muzellec@brown.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

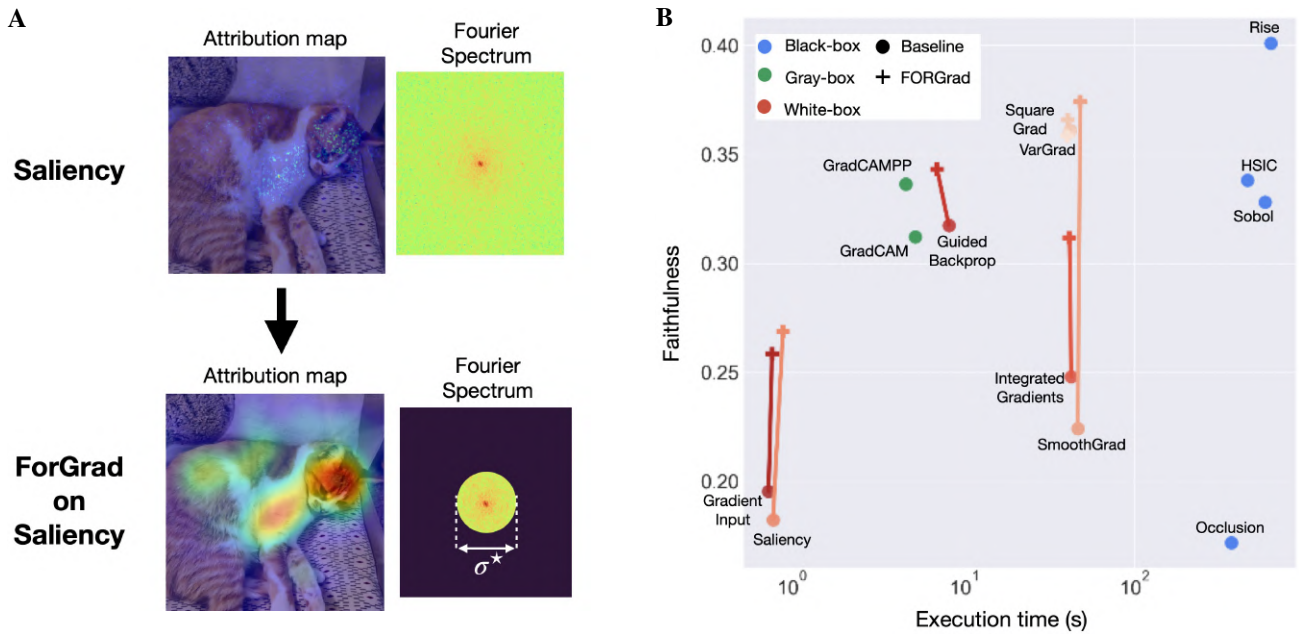

Figure 1: **The FORGrad method. A: FORGrad** estimates an optimal cut-off frequency ($\sigma^\star$) for individual combinations of attribution method and network architecture. It is then used to low-pass filter the gradient signal. **FORGrad** leads to quantitatively better attribution maps. **B: FORGrad** is applicable to all white-box methods (reddish data points), and it consistently improves their faithfulness (see reddish crosses). The x-axis represents the execution time in seconds for each method computed on 100 images. We refer the reader to section 3 for more details on the faithfulness metric.

Consistent with findings from multiple studies (Springenberg et al., 2014; Selvaraju et al., 2017b; Ghorbani et al., 2017; Ancona et al., 2018; Balduzzi et al., 2017; Adebayo et al., 2018; Kim et al., 2019; Tomsett et al., 2019; Rao et al., 2022), we have found empirically that white-box methods typically produce attribution maps with increased power in the high frequencies relative to their black-box counterparts. Our analysis further reveals that these high frequencies are inherited from the gradient itself. We have also found that these high frequencies often arise due to some downsampling operations within the network, such as max-pooling or striding. This observation suggests that white-box methods could be "repaired" by filtering out the gradient signal with carefully chosen network-specific cut-off frequencies.

Here, we introduce **FORGrad** FOurier Reparation of the Gradients, a simple method applicable to all white-box methods with negligible computational overhead. **FORGrad** filters high-frequencies in the gradient using a cut-off frequency adapted for the network architecture and the attribution method used. This frequency is precomputed on training images to maximize the faithfulness of individual attribution methods. Experimentally, we demonstrate that **FORGrad** consistently improves all existing white-box attribution methods. The method and its effects are summarized in Fig. 1.

Overall, **FORGrad** allows white-box methods to compete with the computationally demanding black-box methods on XAI benchmarks. Our contributions are summarized as follows:

- We found that a major source of high-frequency artifacts in attribution maps computed with white-box methods is inherited from the model gradient. These artifacts are responsible for the lower explainability scores of these methods.
- These gradient artifacts are a consequence of the max-pooling and striding operations used in convolutional neural networks (CNNs).
- We introduce **FORGrad** to repair white-box methods. **FORGrad** filters-out frequencies above a certain cut-off value, which is optimized for a given network architecture. **FORGrad** systematically improves the explainability score of white-box methods to levels where they now compete with the computationally intensive black-box methods.

2. Related Work

Black-box attribution methods: In black-box attribution methods, the analytical form and potential internal states of the model are unknown. Therefore, those methods exclusively rely on the network's predictions. The first method, Occlusion (Zeiler & Fergus, 2014), masks individual image regions, one at a time, using an occluding mask set to a baseline value. The corresponding prediction scores are assigned to all pixels within the occluded region, providing an easily interpretable explanation. However, occlusion falls short in capturing the (higher-order) interactions among various image regions – also called coalitions (Idrissi et al., 2023). Sobol (Fel et al., 2021a), along with related methods such as LIME (Ribeiro et al., 2016) and RISE (Petsiuk et al., 2018), address this problem by randomly perturbing multiple regions of the input image simultaneously. Surprisingly, methods like RISE (Petsiuk et al., 2018) Sobol (Fel et al., 2021a) and HSIC (Novello et al., 2022) have outperformed their white-box counterparts which have full access to network inner information (see Fig. 1B and Fig. 3B). However, they come with a trade-off in terms of speed: black-box methods typically need thousands of network queries to generate a comprehensive explanation (see Fig. 3B).

White-box attribution methods: In contrast to black-box methods, white-box methods have full access to the network information (e.g., the weights, the architecture) and rely on finer information contained in the gradient signal. It was the original idea of the Saliency method (Simonyan et al., 2013), the first to back-propagate the gradient from the output to the network's input, indicating which pixels affect the decision score the most. Although computationally very efficient, Saliency is limited because it focuses on the influence of individual pixels in an infinitesimal neighborhood in the input image. For instance, it has been shown that gradients often vanish when the prediction score to be explained is near the maximum value (Sundararajan et al., 2017; Miglani et al., 2020). Integrated Gradient (Sundararajan et al., 2017), SmoothGrad (Smilkov et al., 2017), and SquareGrad (Hooker et al., 2019; Seo et al., 2018) partially address this issue by accumulating gradients. However, such white-box methods still lag behind black-box attribution methods.

Gray-box attribution methods: The distinction between white-box and black-box methods is nuanced, particularly as some approaches leverage only partial knowledge of the network. Grad-CAM (Selvaraju et al., 2017a) and Grad-CAM++ (Chattopadhay et al., 2018) fall within this category. These methods use the gradient information only of the last layers of the network, as opposed to the full back-propagation done by white-box methods. Specifically, Grad-CAM and Grad-CAM++ generate visual explanations for class labels using a weighted combination of the positive partial derivatives of the last convolutional layer feature maps with respect to a specific class score. This process generates small sample-specific maps that are then upsampled to fit the input size. Although these methods perform relatively well and are computationally efficient, the corresponding attribution maps often lack fine details (Nam et al., 2021). This limitation primarily stems from the low-spatial resolution gradient maps they rely on to produce attribution maps.

3. Notations, Metrics, and Networks

Notations: We consider a general supervised learning setting, where a classifier $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$ maps images from an input space $\mathcal{X} \subseteq \mathbb{R}^{W \times H}$ to an output space $\mathcal{Y} \subseteq \mathbb{R}$. We denote $\mathcal{F}$ the Discrete Fourier Transform (DFT) on $\mathbb{R}^{W \times H}$. Importantly, all Fourier spectra shown in this article are symmetrized, i.e., we always shift the low-frequency components to the center of the spectrum. We recall that an attribution method is a function $\boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^{W \times H}$ that maps an input of interest to its corresponding importance scores $\boldsymbol{\varphi}(\boldsymbol{x})$. Finally, we denote by $\boldsymbol{\varphi}_\sigma(\boldsymbol{x})$ the attribution method in which high frequencies are filtered out using a cut-off value of σ.

Fourier signature and power/frequency slope: In this article, we analyze the Fourier signature of attribution maps produced by white-box and black-box methods. To do so, we compute the amplitude of the Fourier spectrum, $|(\mathcal{F} \circ \boldsymbol{\varphi})(\boldsymbol{x})|$ for attribution maps obtained with different methods. We average the spectrum over $1,000$ images. In Fig. 2, we illustrate how to extract the average Fourier signature from the Fourier spectrum computed over attribution (or gradient) maps. This signature is computed using the circular average of the Fourier amplitudes (averaged over all θ values) at each frequency (i.e., each R). We then derive a single scalar, called the power/frequency slope, as the slope of the best linear fit of the Fourier signature. Intuitively, a steeply negative slope indicates a low power in the high-frequencies. Inversely, a slope close to zero suggests that the underlying attribution maps contain a lot of high-frequency signals.

XAI metrics: To evaluate XAI performance, we use the faithfulness metric, denoted $F(\boldsymbol{x}, \boldsymbol{\varphi})$, which is a measure that includes the Insertion and Deletion metrics (Petsiuk et al., 2018). Those metrics are widely used in XAI to evaluate the quality of attribution methods (Fong & Vedaldi, 2017). The Deletion metric, $D(\boldsymbol{x}, \boldsymbol{\varphi})$, measures the evolution of the prediction probability when one incrementally removes pixels by replacing them with a baseline value according to their attribution score. Insertion, $I(\boldsymbol{x}, \boldsymbol{\varphi})$, inserts pixels gradually into a baseline input. The faithfulness is obtained using $F(\boldsymbol{x}, \boldsymbol{\varphi}) = I(\boldsymbol{x}, \boldsymbol{\varphi}) - D(\boldsymbol{x}, \boldsymbol{\varphi})$. Note that the faithfulness metric has shown to be effective in evaluating attribution methods (Samek et al., 2016; Li et al., 2022).

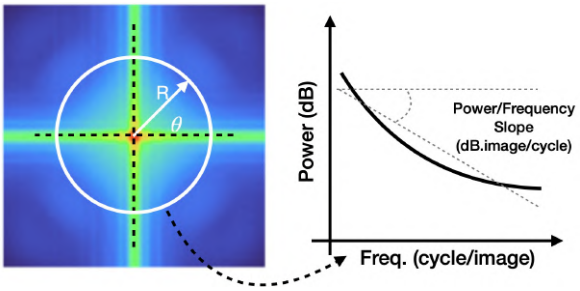

Figure 2: **Fourier signature and power/frequency slope.** The Fourier signature is computed using a circular average of the Fourier amplitudes (averaged over all θ values) at each frequency (i.e., each R). The power/frequency slope summarizes in one scalar the Fourier signature.

Additionally, we also evaluate attribution methods using the μFidelity, denoted μF (Bhatt et al., 2020). The fidelity metric serves to verify the correlation between the attribution score and a random subset of pixels. To achieve this, a set of pixels is randomly chosen and set to a baseline state, after which a prediction score is obtained. The fidelity metric evaluates the correlation between the decrease in the score and the significance of the explanation for each random subset created. Finally, the sensitivity metric, $S(\boldsymbol{x}, \boldsymbol{\varphi})$ (Bhatt et al., 2020), measures a method's stability by calculating the distance between two explanations. These explanations are generated by the same method, with the first using the original input and the second using the same input perturbed with random noise. A good method, according to this metric, should produce similar explanations for the two inputs, therefore obtaining a low sensitivity score.

Networks: We analyze attribution methods on three different representative ImageNet-trained classifiers: ResNet50V2 (He et al., 2016a), ConvNeXt (Liu et al., 2022), and ViT (Dosovitskiy et al., 2020). For concision, we include in the main text only the results obtained with the ResNet50V2 network and we refer the reader to supplementary information for the 2 other networks. All the results obtained on ConvNeXt and ViT are consistent with those of ResNet50V2.

4. White-box Methods are Contaminated by High-Frequencies from the Gradient

4.1. Comparison between White-Box and Black-Box Methods

White-box methods are contaminated with high-frequency signal: Fig. 3A shows the average Fourier signature for white-box and black-box attribution methods (see Fig. A.1 for the 2D visualization). We observe that white-box methods produce attribution maps with a greater high-frequency power compared to black-box attribution maps (red curves vs. blue curves in Fig. 3A). The gray-box methods lie in between white-box and black-box methods.

White-box methods are computationally more efficient but have lower faithfulness: We quantify the computational time of each method to generate an explanation on 100 ImageNet images with an Nvidia T4. The computational time is reported for each attribution method using numbers placed under their corresponding bar in Fig. 3B. We refer the reader to section 3 for more details on the faithfulness computation. Overall, Fig. 3B demonstrates that white-box (and gray-box) methods are computationally more efficient than black-box methods. In addition, it shows that white-box methods have a lower power/frequency slope and a lower faithfulness than black-box methods. This suggests that the presence of high-frequency artifacts in white-box methods might be a limiting factor in achieving XAI performance similar to black-box methods.

4.2. A Low-Pass Filtered Gradient Approximates Well the Original Gradient

White-box methods produce attribution maps by relying on the gradients of the classifier w.r.t. the input, denoted $\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x})$. The prevalence of high frequencies in these attribution maps suggests that the gradients harbor high-frequency artifacts. To study such artifacts in the gradients, we leverage the first-order approximation of the model, i.e., $\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{\varepsilon}) \approx \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\varepsilon} \nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x})$, and we compute the ℓ_2 approximation error, over different values of $\boldsymbol{\varepsilon}$.

We characterize the error between the Taylor expansion and the actual function value for a specified cut-off frequency of the gradient, denoted as σ, through the following definition: $\zeta(\boldsymbol{x}, \sigma) = ||\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{\varepsilon}) - (\boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\varepsilon} \nabla_{\sigma} \boldsymbol{f}(\boldsymbol{x}))||_2$. We then define our metric as $\zeta(\boldsymbol{x}, \sigma)/\zeta(\boldsymbol{x}, \sigma_{\text{max}})$, with $\sigma_{\text{max}} = 224$. Here, $\nabla_{\sigma} \boldsymbol{f}(\boldsymbol{x})$ represents the gradient at $\boldsymbol{x}$ with high frequencies removed, up to the cut-off frequency σ. We compute the error ζ for various filtered gradients with a cut-off frequency $\sigma \in \{224, \ldots, 5\}$ (where a high σ represents a very minor filtering and low σ implies that only the very low frequencies are kept), as well as three control conditions representing the absence of information in the gradient.

In Fig. 4.A, the blueish curves show the evolution of the approximation error from the baseline (gradient not filtered – most green curve) to the most filtered gradient ($\sigma = 5$ – most blue curve). We additionally report the approximation for three control conditions (zero gradients, permuted gradients, and uniform noise gradients). In Fig. 4.A, we observe that low-pass filtering gradients elicit a ℓ_2 approximation error very close to the baseline. These results suggest that the filtered gradient, which remains closer to the first-order term than any of the control conditions, still approximates the non-filtered gradient well when defined as the first-order

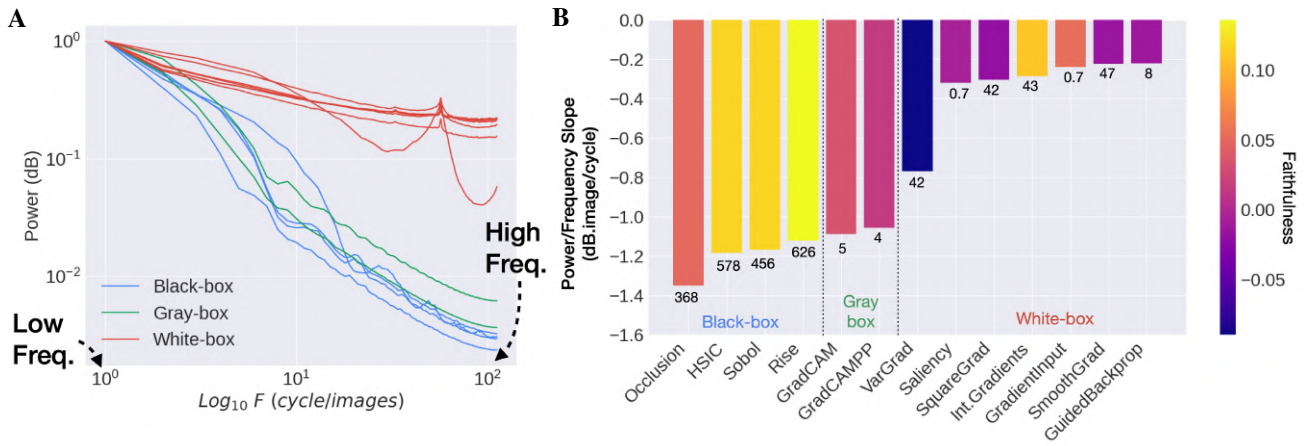

Figure 3: **High-frequency power in attribution methods. A:** Fourier signature of white-box and black-box attribution methods. White-box methods produce attribution maps with increased power in the high frequencies. **B:** Faithfulness and power/frequency slope for several attribution methods. The numbers below the bar correspond to this method's computational time, measured using a set of 100 images (ImageNet) with an Nvidia T4. White-box methods show lower faithfulness but significantly better computational complexity compared to black-box methods. See section 3 for more details on how faithfulness, the power/frequency slope, and the Fourier signature were computed.

term of a Taylor expansion. Given the linear nature of the first-order approximation, one might consider that the first-order gradient mostly reflects low-frequency content. To control for this, we have included a condition featuring synthetic low-frequency gradients (in the form of zeros everywhere). Such a condition elicits high approximation error (see red curve in Fig. 4.A), suggesting that the first-order approximation of the gradient does not capture only low-frequencies. We refer the reader to Fig. A.2 for similar results on ConvNeXt and ViT, as well as an additional control condition representing non-informative low-frequency content under the form of a 2D Gaussian.

4.3. High-Frequency Artifacts Stems from Max-Pooling Operations

Our investigation now turns to understanding why high-frequency artifacts appear in gradients. In Fig. 4B, we analyze the power spectrum of various gradients in a ResNet50, along with their visualization. Consistent with findings from Olah et al. (2017), we observe a clear checkerboard pattern in the gradients following a max-pooling. This pattern is reduced after an average-pooling due to a different mechanism in the backward implementation of both operations (see Fig. A.3 for more visualizations). Their respective Fourier signature (Fig. 4B right) confirm our hypothesis: the max-pooling operation shows more power in the high frequencies compared to the average-pooling (see Fig. A.4 for all pooling layers). We additionally investigate the contribution of every layer in the frequency power of the gradients. We analyze the power/frequency slope of gradients across various layers of ResNet50 (see Fig. A.5). We observe a jump in the power/frequency slope at each max-pooling operation in between ResNet blocks (see blue curve in Fig. A.5). To discern whether this increase in high frequency is a result of the network architecture or the training process, we conduct a similar analysis with an untrained ResNet50 (black curve in Fig.A.5) and observe a similar trend. Subsequently, we replace max-pooling with average-pooling to ensure the preservation of information continuity in the gradient (red curve in Fig.A.5). We observe a clear divergence in the power/frequency slope between max-pooling and average-pooling networks, highlighting the crucial role of max-pooling in creating high-frequency signals in gradients. We observe a similar pattern of results with a VGG16 (Simonyan & Zisserman, 2014) architecture (see Figures A.6,A.7, A.8). Additionally, as noted by Odena et al. (2016), this effect can be mitigated by adjusting the stride value during downsampling. Consequently, in Fig. A.9, we examine the impact of striding on high-frequency power in a VGG16 network.

In summary, we have established that the high-frequency artifacts contaminating white-box methods originate from the gradient. Specifically, max-pooling operations within the network appear to be a primary source of these high-frequency artifacts. These insights underscore the potential benefits of implementing a low-pass filter with a cut-off frequency adapted to the network architecture to repair white-box attribution methods.

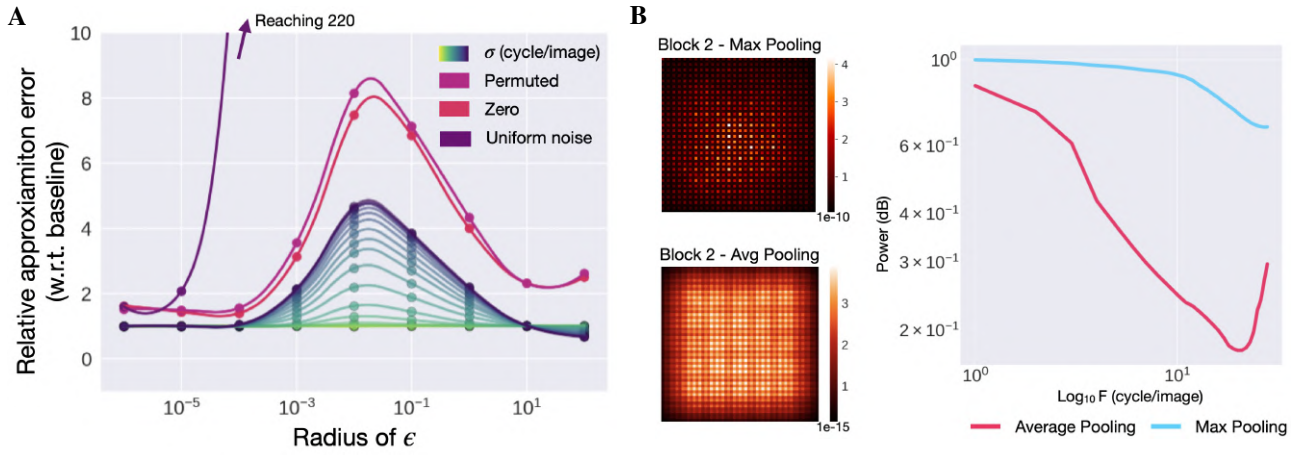

Figure 4: **Gradients are contaminated with high frequencies and stem from MaxPooling operations**. **A:** We evaluate the importance of high-frequency content in the gradient using a first-order approximation of the model, i.e., $\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\varepsilon}) \approx \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\varepsilon}\nabla_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})$, and we compute the ℓ_2 approximation error when we remove high frequency up to σ: $\zeta(\boldsymbol{x}, \sigma)$ relative to the error when we do not filter $\zeta(\boldsymbol{x}, \sigma_{\text{max}})$. Gradients in which we remove high-frequency content (dark blue) produce an error closer to the baseline (green) compared to the control conditions (pink). **B:** Example of gradients in a ResNet architecture after a pooling operation, along with their Fourier signature. MaxPooling operations elicit high-frequency power in the gradient.

5. FORGrad: FOurier Reparation of the Gradients

With **FORGrad** we propose to filter out high-frequency content that is considered noise to improve the XAI performance of white-box attribution methods. We therefore apply a low-pass filter on the Fourier spectrum of the gradient of the classifier w.r.t the input (i.e. $\nabla_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})$). This low-pass filter is equipped with a mechanism to adjust the cut-off frequency, $\sigma^\star$, based on the specific network architecture and attribution method.

An adapted $\sigma^\star$ per model and attribution method: To identify the ideal cut-off frequency, $\sigma^\star$, we introduce a heuristic to maximize attribution faithfulness. For each network and attribution method, we first assess the relationship between faithfulness and different cut-off frequencies, σ in cycle/image, as shown in Fig.5B (see Fig. A.10 for the sigma selection per model and method). The faithfulness is evaluated on $1,280$ images from the ImageNet validation set (denoted $\mathcal{V}$) for a spectrum of cut-off frequencies. These frequencies range from 224, indicative of an unfiltered gradient, to 0, representing a fully filtered gradient. Detailed information about the faithfulness metric can be found in section 3. Here, we select the optimal bandwidth $\sigma^\star$ such that $\sigma^\star = \arg\max_\sigma \mathbb{E}_{\boldsymbol{x}\sim\mathcal{V}} F(\boldsymbol{\varphi}_\sigma(\boldsymbol{x}))$.

Fig. 5B illustrates the sensitivity of the faithfulness to changes in the cut-off frequency, σ, underscoring the necessity of selecting an optimal $\sigma^\star$. In Fig. 5A, we provide visual examples using the Gradient Input method (Ancona et al., 2018), where the gradient is low-pass filtered at various cut-off frequencies. One can observe that the initial noisy patterns (Fig. 5A, $\sigma = 220$) progressively evolve into clearer and larger patches (Fig. 5A, $\sigma = 80$), culminating in saliency maps that effectively emphasize the critical features for object recognition (Fig. 5A, $\sigma = 10$). However, at very low σ values (Fig. 5A, $\sigma = 5$), the attribution map becomes overly diffuse, reducing interpretability. Fig. 5D compares the optimal $\sigma^\star$ values across different ResNetV2 (He et al., 2016b) architectures and attribution methods. We observe that even with the same attribution method, distinct architectures achieve their maximum faithfulness at different $\sigma^\star$ values. Likewise, the same network architecture can yield different optimal $\sigma^\star$ values when leveraging different attribution methods. This variability highlights the crucial role of a specific cut-off frequency, emphasizing that a one-size-fits-all approach may not be effective in maximizing the faithfulness of attribution methods in different contexts. Additionally, Fig. 5C presents a side-by-side comparison of attribution maps generated using the SmoothGrad method (Smilkov et al., 2017) (top-row) versus those refined with the **FORGrad** method (bottom row). A qualitative analysis shows that the **FORGrad** method yields more interpretable explanations, with a sharper focus on crucial features for object recognition. We refer the reader to Fig. A.12 for more examples of all white-box methods.

FORGrad allows white-box methods to rival with black-box methods: We now evaluate the **FORGrad** method using XAI benchmarks, focusing on three established met-

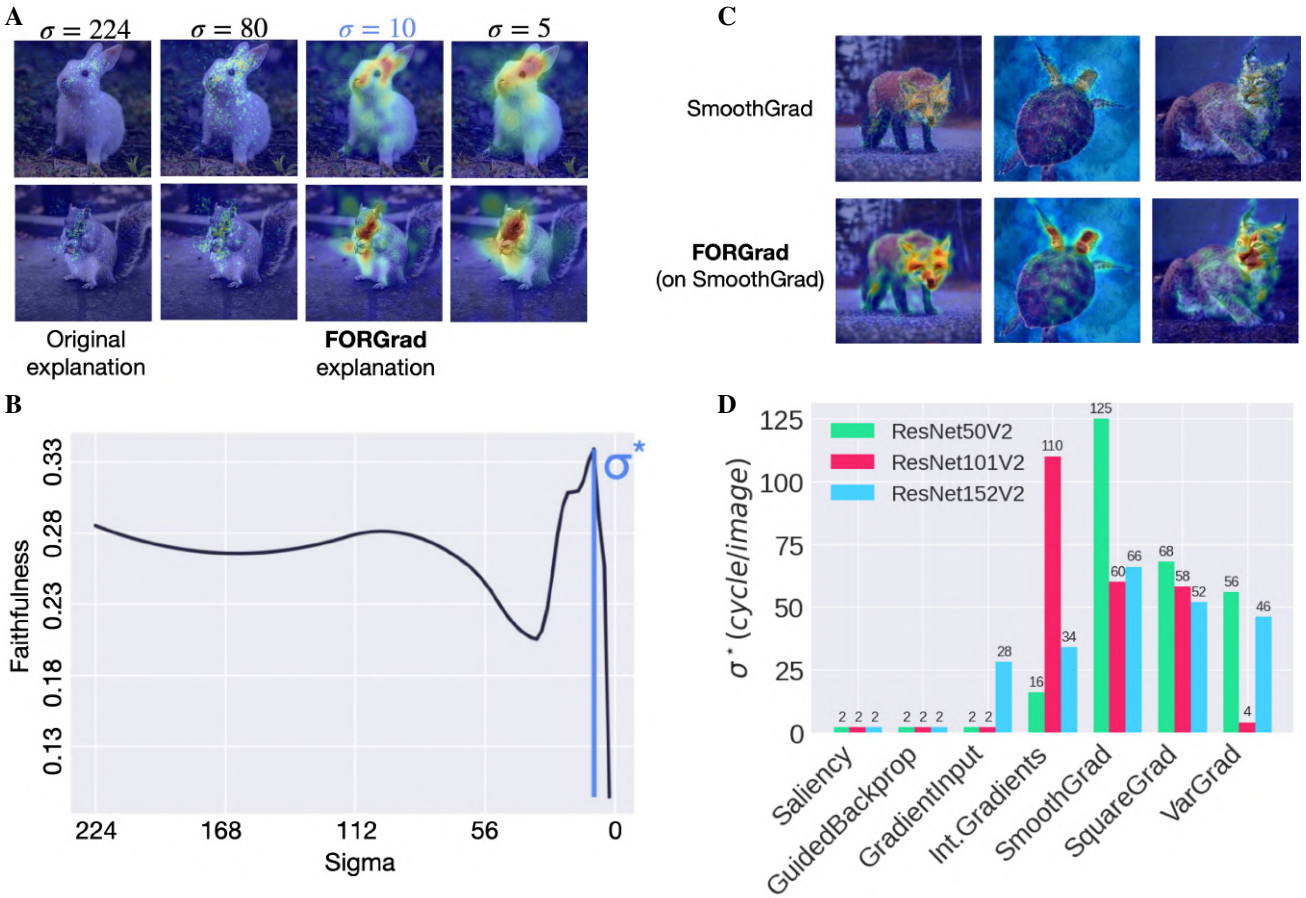

Figure 5: **The FORGrad method**. **A:** Visual examples in which the gradient is low-pass filtered at various cut-off frequencies. Without any filtering ($\sigma = 224$), the attribution maps are similar to those obtained with original white-box methods (i.e., Gradient-Input here). **FORGrad** finds an optimal cut-off frequency at $\sigma = 10$. **B:** Sensitivity of the faithfulness to changes in the cut-off frequency. The faithfulness is evaluated on 1,280 images from the ImageNet validation set (see section 3 for more details). The frequencies range from 224, indicative of an unfiltered gradient, to 0, representing a fully filtered gradient. **C:** Side-by-side comparison of attribution maps generated using the SmoothGrad method (top-row) versus those refined with the **FORGrad** method (bottom row). **D:** Evolution of the optimal $\sigma^\star$ values across different ResNetV2 architectures and attribution methods. The variability highlights the crucial role of a tailored cut-off frequency.

rics: Faithfulness, μFidelity, and Sensitivity (detailed in section 3). Our findings, presented in Table 1, compare the original performance of white-box attribution methods, their enhanced versions via the **FORGrad** method (marked with a $\star$), black-box, and gray-box methods. This comparison also includes the computational time measured on 100 images with an Nvidia T4 required to produce attribution maps. We showcase results on a ResNet50 network (He et al., 2016a) in Table 1, and provide additional results for ConvNext (Liu et al., 2022) and ViT (Dosovitskiy et al., 2020) in Table 4. All metrics are computed on a subset of the ImageNet validation set made of 1280 images. Importantly, this subset does not include the images used to compute $\sigma^\star$. Three key observations emerge from Table 1. First, **FORGrad** consistently boosts the faithfulness metrics for white-box methods (the quantitative improvement is validated by a statistical effect given by a Bayesian Anova analysis – section A.6). This observation experimentally validates the **FORGrad** method. Second, the μFidelity metric shows little variations for most of the **FORGrad** white-box methods. We attribute this observation to the fact that the μFidelity metric is strongly biased toward high-frequency content: random attribution maps with higher-frequency contents exhibit higher μFidelity (see Table 3 and Fig. A.13). However, the sensitivity metric, similarly to the faithfulness, is systematically improved for most models and methods, despite being evaluated with a $\sigma^\star$ optimized on the faithfulness. Third, in terms of computations, white-box (and gray-box) methods outperform black-box approaches. This increased efficiency stems from white-box attribution methods using more information per network pass, thereby reducing the need for multiple network queries. Additionally, we validate

		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

Table 1: **Results on Faithfulness and Stability metrics**. Faithfulness, Fidelity, and Stability scores were obtained on 1,280 ImageNet validation set images, on an Nvidia T4 (For Faithfulness and Fidelity, higher is better). Time in seconds corresponds to the generation of 100 (ImageNet) explanations on an Nvidia T4. The first and second best results are in **bold** and <u>underlined</u>.

the optimality of our method by computing **FORGrad** on the attribution maps instead of the gradients. In Table 7, we compare the non-filtered metrics with the ones obtained with **FORGrad** on the attribution maps and **FORGrad** on the gradients (before computing the explanation). We observe a systematic improvement of the scores with **FORGrad** but more significant when applied to the gradients. Finally, we ensure that **FORGrad** does not affect the validity of white-box methods according to Adebayo et al. (2018) sanity check. We show in Fig. A.14 the resulting maps after randomizing the weights of a ResNet50 on white-box methods combined with **FORGrad**. The methods already failing the test before **FORGrad** still do, while the successful ones continue to pass the sanity check. All in all, **FORGrad** discards high-frequency artifacts to enhance low-frequency content of the attribution maps provided by the white-box methods by filtering out the noisy high-frequencies in the gradient, making such methods rival with low-frequency black-box methods, as summarized in Fig. A.15.

In Table 2, we aim to find the best attribution methods for ResNet50. To do so, we synthesize the faithfulness, μfidelity, and sensitivity metrics into one single, comprehensive score. This score, labeled $(F + \mu F) - S$, corresponds to the sum of the faithfulness and the μfidelity, minus the sensitivity. Table 2 reveals a notable shift in method rankings when integrating **FORGrad**. This change results in

	ResNet50	
	Original	**FORGrad**
1	VarGrad (0.82)	**SquareGrad⋆ (0.82)**
2	SquareGrad (0.81)	**VarGrad⋆ (0.82)**
3	GuidedBackprop (0.48)	**SmoothGrad⋆ (0.60)**
4	SmoothGrad (0.46)	**GuidedBackprop⋆ (0.58)**
5	GradCAM++ (0.42)	GradCAM (0.42)

Table 2: **Global ranking before (original) and after FORGrad.** For each model, we show the five attribution methods with the highest metrics, before and after applying **FORGrad**. The explanation maps were computed on 1280 images from the validation set of ImageNet, based on an aggregation of the three metrics computed by $(F(\boldsymbol{x}, \boldsymbol{\varphi}) + \mu F(\boldsymbol{x}, \boldsymbol{\varphi}) - S(\boldsymbol{x}, \boldsymbol{\varphi})$.

at least 4 white-box methods in the top-5 position for the ResNet50 network. Similar findings for ConvNext and ViT are detailed in Table 5.

In conclusion, the **FORGrad** method not only enhances the performance of white-box methods to rival or even surpass black-box methods but also achieves this with greater computational efficiency.

6. Conclusion and Perspectives

This research started with the observation that white-box and black-box attribution methods produce attribution maps with different power spectra. Specifically, white-box methods show greater high-frequency power. We found that this high-frequency content originates from the gradient information used by white-box methods. We demonstrate that selectively filtering out these high frequencies does not alter the approximation error of the original gradient (using a first-order approximation). It suggests that such high-frequency content is non-informative and can be considered as noise. Further analysis revealed a strong connection between high frequencies in gradients and certain architectural constraints, including max-pooling operations. This led us to hypothesize that filtering out these high frequencies would enhance white-box attribution methods that rely on gradient information. To address this, we developed **FORGrad** a low-pass filter with an optimal cut-off frequency tailored to each model architecture and attribution method. The application of **FORGrad** resulted in substantial improvements in XAI performance across all white-box attribution methods, challenging the traditional hierarchy of white-box and black-box methods. Notably, with **FORGrad** white-box methods now dominate the top-5 rankings while also being more computationally efficient than their black-box counterparts.

Our current results demonstrate the efficiency of **FORGrad** on gradients containing artifacts. For this reason, the more high frequency is contained in the gradient, the

more striking the improvement with **FORGrad** will be. Conversely, if filtering does not ameliorate the scores, **FORGrad** will not filter out anything, resulting in the original unfiltered explanation. Such a case is likely to happen with networks containing smooth gradients by design, such as 1-Lipschitz networks (Fel et al., 2022).

Our finding that the max-pooling operations significantly amplify high frequency in the gradient w.r.t input is surprising. This observation opens the doors to future research focused on designing novel pooling techniques that reduce high frequencies in the gradient itself. Consistent with these results, the impact of **FORGrad** on ViT is less striking than on the Convolution model. This implies that a different phenomenon could happen in Transformer architectures and an investigation could lead to a better interpretability of this family of networks. Additionally, conducting a systematic analysis of the Fourier spectrum of other architectural components, such as the normalization or attention layers, could lead to further improvements. Ideally, such an exploration would result in a comprehensive list of each architectural element's contribution to the decision process. Studying their respective impacts would contribute to building neural networks that are interpretable by design.

The impact of **FORGrad** is measured by the standard metrics of XAI. We evaluated **FORGrad** on three different metrics and observed a clear improvement in two of them. However, finding appropriate metrics is still an open question in this field, and these results could be affected by the development of new metrics. In Bhatt et al. (2020), the authors advocate for a way to combine several XAI metrics to obtain the best method. They propose an aggregate of Faithfulness, Sensitivity, and Complexity of the explanation, arguing that the best explanation should maximize the first and minimize the latter two. Here, we evaluate **FORGrad** on two of these metrics and complete this framework with the Computational cost measured by the time in seconds. We, therefore, believe that **FORGrad** remains an optimal solution to compute faithful explanations in a reasonable amount of time, given the current metrics to evaluate the quality of an explanation. More interestingly, one possible explanation for why black-box methods tend to outperform white-box methods is that their optimization objectives align with the evaluation metrics (e.g., Occlusion with Deletion metric). In contrast, the gradients result from optimizing the model for a classification task. Therefore, it is encouraging to achieve scores comparable to those of black-box methods using white-box attribution methods with **FORGrad** bringing Saliency and other gradient-based methods back into prominence.

So far, we have focused on gradient filtering to improve white-box attribution methods. One promising avenue would be to filter out high-frequencies in the gradient during training to reduce the presence of noisy artifacts during the optimization process and inherently improve its quality. We hypothesize that this approach may lead to better performance and faster convergence, as well as more qualitative information to explain the model using white-box methods. Networks equipped with such a filtering mechanism may also demonstrate enhanced adversarial robustness. We hope our work will spur interest in studying network frequency profiles, as it could lead to major improvements in terms of interpretability and robustness. Overall, we anticipate that our research will foster broader adoption of simpler and more efficient white-box methods for explainability, offering a better balance between faithfulness and computational efficiency.

Acknowledgments

This work was supported by grants from ONR (N00014-19-1-2029) and NSF (IIS-1912280 and EAR-1925481) to TS and RV and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004). The Carney Institute for Brain Science and the Center for Computation and Visualization (CCV) provided additional support. We acknowledge the Cloud TPU hardware resources that Google made available via the TensorFlow Research Cloud (TFRC) program and computing hardware supported by NIH Office of the Director grant S10OD025181.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Agarwal, C., Johnson, N., Pawelczyk, M., Krishna, S., Saxena, E., Zitnik, M., and Lakkaraju, H. Rethinking stability for attribution-based explanations. In *ICLR 2022 Workshop on PAIR 2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022.

Agarwal, S., Jabbari, S., Agarwal, C., Upadhyay, S., Wu, S., and Lakkaraju, H. Towards the unification and robustness of perturbation and gradient based explanations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Balduzzi, D., Frean, M., Leary, L., Lewis, J., Ma, K. W.-D., and McWilliams, B. The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning*, pp. 342–350. PMLR, 2017.

Bhatt, U., Weller, A., and Moura, J. M. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2018.

Colin, J., Fel, T., Cadène, R., and Serre, T. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017a.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *ArXiv e-print*, 2017b.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fel, T., Cadène, R., Chalvidal, M., Cord, M., Vigouroux, D., and Serre, T. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. *Advances in Neural Information Processing Systems*, 34: 26005–26014, 2021a.

Fel, T., Cadene, R., Chalvidal, M., Cord, M., Vigouroux, D., and Serre, T. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Fel, T., Vigouroux, D., Cadène, R., and Serre, T. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 720–730, 2022.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Han, T., Srinivas, S., and Lakkaraju, H. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Hase, P. and Bansal, M. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 630–645. Springer, 2016b.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Idrissi, M. I., Bousquet, N., Gamboa, F., Iooss, B., and Loubes, J.-M. On the coalitional decomposition of parameters of interest, 2023.

Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 624–635, 2021.

Kim, B., Seo, J., Jeon, S., Koo, J., Choe, J., and Jeon, T. Why are saliency maps noisy? cause of and solution to noisy saliency maps. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4149–4157. IEEE, 2019.

Kim, S. S. Y., Meister, N., Ramaswamy, V. V., Fong, R., and Russakovsky, O. HIVE: Evaluating the human interpretability of visual explanations. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2022.

Li, H., Li, Z., Ma, R., and Wu, T. Fd-cam: Improving faithfulness and discriminability of visual explanation for cnns. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 1300–1306. IEEE, 2022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, June 2022.

Masson, M. E. A tutorial on a practical bayesian alternative to null-hypothesis significance testing. *Behavior research methods*, 43:679–690, 2011.

Miglani, V., Kokhlikyan, N., Alsallakh, B., Martin, M., and Reblitz-Richardson, O. Investigating saturation effects in integrated gradients. 2020.

Nam, W.-J., Choi, J., and Lee, S.-W. Interpreting deep neural networks with relative sectional propagation by analyzing comparative gradients and hostile activations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11604–11612, 2021.

Nguyen, G., Kim, D., and Nguyen, A. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Novello, P., Fel, T., and Vigouroux, D. Making sense of dependence: Efficient black-box explanations using dependence measure. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Odena, A., Dumoulin, V., and Olah, C. Deconvolution and checkerboard artifacts. *Distill*, 2016. doi: 10.23915/distill.00003. URL http://distill.pub/2016/deconv-checkerboard.

Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Petsiuk, V., Das, A., and Saenko, K. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

Pukdee, R., Sam, D., Kolter, J. Z., Balcan, M.-F., and Ravikumar, P. Learning with explanation constraints. *arXiv preprint arXiv:2303.14496*, 2023.

Rao, S., Böhle, M., and Schiele, B. Towards better understanding attribution methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, 2016.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 2016.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017a.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017b.

Seo, J., Choe, J., Koo, J., Jeon, S., Kim, B., and Jeon, T. Noise-adding methods of saliency map as series of higher order partial derivative. In *Workshop on Human Interpretability in Machine Learning, Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Serrurier, M., Mamalet, F., Fel, T., Béthune, L., and Boissin, T. When adversarial attacks become interpretable counterfactual explanations. *arXiv preprint arXiv:2206.06854*, 2022.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop, Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

Sixt, L., Granz, M., and Landgraf, T. When explanations lie: Why many modified bp attributions fail. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Slack, D., Hilgard, A., Singh, S., and Lakkaraju, H. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. In *Workshop on Visualization for Deep Learning, Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., and Preece, A. Sanity checks for saliency metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2014.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

A. Appendix

A.1. Additional results

Frequency power per category. Our quantitative measure of high-frequency power derives from the Fourier signature (the 1D Fourier spectrum). The spectra per method, as shown in Fig. 3, show a clear difference between categories. As these spectra are derived from their 2D equivalent, we provide in Fig. A.1, the 2D visualization. These spectra highlight the same information: white-box methods show a Fourier spectrum with a lot of power in the high-frequency, conversely to gray and black-box methods.

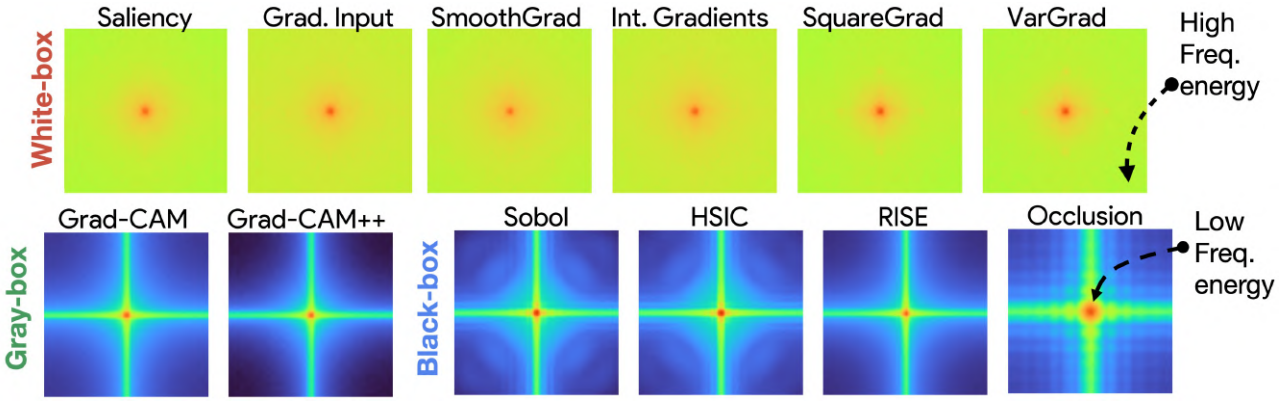

Figure A.1: **2D Fourier spectrum of attribution methods.** We show on the top row the Fourier spectrum of white-box attribution methods and of the gray and black-box methods on the bottom row, computed with a ResNet50V2 (each spectrum is normalized to highlight values ranging within the same boundaries). The three families can be distinguished by methods but also by their signature in the Fourier domain. The former method has magnitudes largely concentrated in the low frequencies, while the latter is more spread out: it features non-trivial magnitudes almost everywhere, including in high frequencies. (Blue = low power, Green = high power, Red = very high power)

A.2. Gradient approximation with ConvNeXt and ViT

In this section, we show the results of the gradient approximation presented in the main text on ConvNeXt and ViT. For each network, we evaluate the importance of high-frequency content in the gradient using a first-order approximation of the model, i.e. $\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\varepsilon}) \approx \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\varepsilon}\nabla_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})$, and we compute the ℓ_2 approximation error, i.e. $\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\varepsilon}) - (\boldsymbol{f}(\boldsymbol{x}) - \epsilon.\nabla_{\sigma})\|_2 / \|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\varepsilon}) - (\boldsymbol{f}(\boldsymbol{x}) - \epsilon.\nabla_{\sigma_{max}})\|_2$. We additionally provide four control conditions representing the absence of relevant information in the gradient using a gradient containing permuted content, zeros, uniform noise, or random 2D gaussians. The left plot shows the results on ConvNeXt and the right plot for ViT. The reader can notice that compared to Fig. 4.A, the approximation errors of the permuted and zero control conditions reach an order of magnitude more than the most filtered gradient, confirming our precedent results with ResNet50. The 2D Gaussian control condition represents information of low-frequency content under the form of randomly located 2D Gaussians instead of the gradient. Contrary to the "zero" condition (also representing low-frequency information), the activity is not null and might be more informative. However, it reaches the same approximation as the other baseline, being at least twice as high as the more filtered gradient. These additional results confirm the hypothesis that high frequencies in the gradient can be considered not as informative as the low frequencies.

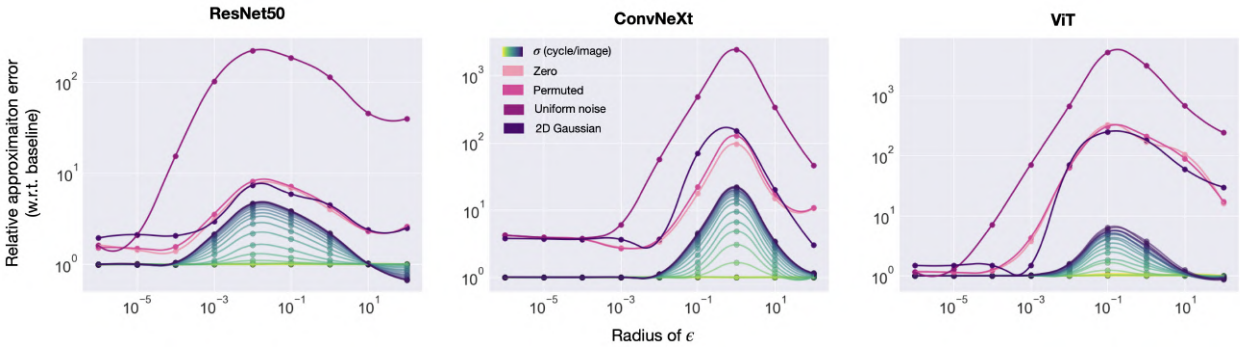

Figure A.2: **Gradient approximation experiment on ResNet50, ConvNeXt and ViT.** We evaluate the importance of high-frequency content in the gradient using a first-order approximation of the model, i.e. $\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\varepsilon}) \approx \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\varepsilon}\nabla_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})$, and we compute the ℓ_2 approximation error, i.e. $\|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\varepsilon}) - (\boldsymbol{f}(\boldsymbol{x}) - \epsilon.\nabla_{\sigma})\|_2 / \|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\varepsilon}) - (\boldsymbol{f}(\boldsymbol{x}) - \epsilon.\nabla_{\sigma_{max}})\|_2$. Gradients in which we remove high-frequency content (dark blue) produce an error closer to the baseline (green), compared to the control conditions (purple).

A.3. Investigating the impact of MaxPooling on high-frequency content

The following section details the role of MaxPooling in introducing high-frequency artifacts in the gradient. We obtain the gradient w.r.t the input by running an inference on a model (pre-trained or not), and back-propagating the information back to the input layer. In the section, we are therefore interested in the direction of the information from the prediction to the input, and we investigate the role of each operation on the content of the gradient.

Downsampling operations in a ResNet50V2. We start by providing additional visualizations of the gradient at several stages of a ResNet50V2 in Fig. A.3. We compare the content of a gradient provided by a pre-trained model on ImageNet, a random model, and a model where the MaxPooling operations as well as started convolutions were replaced by an AveragePooling. The first block, starting from the prediction – and denoted *conv5* here – shows no strong high-frequency artifact. However, after applying a downsampling operation (upsampling on the gradient), we observe some checkerboard-like patterns emerging in the architectures with MaxPooling. The pattern is however minimized in the architecture doted of AveragePooling. The deeper we look into the network, the more high frequencies are present in the gradient.

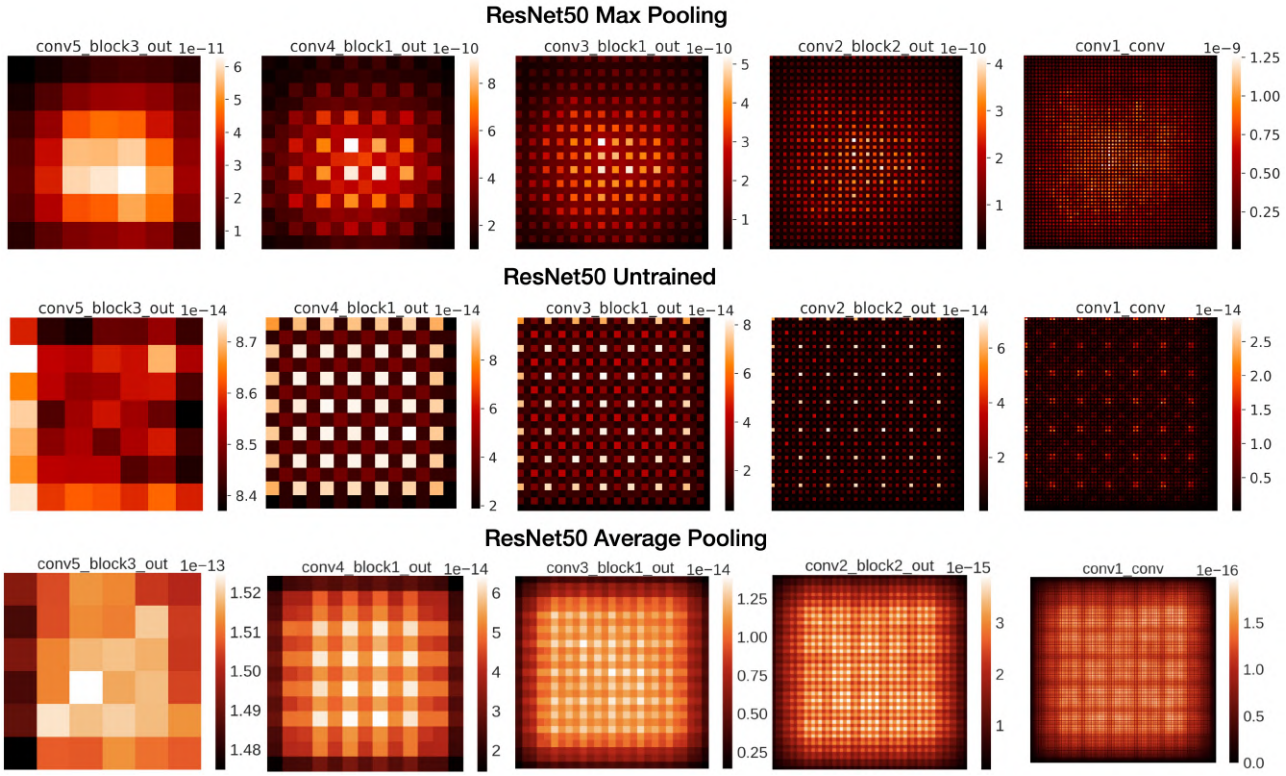

Figure A.3: Visualization of the gradient (w.r.t. the input) probed at different layers. We observe that both in the ResNet MaxPooling and the ResNet untrained, a checkerboard signal emerges in the gradients. Such a signal might cause an increased power in high frequencies. In contrast, this signal is attenuated in the ResNet50 Average Pooling. It suggests that MaxPooling operations, creating such a checkerboard pattern, might contaminate the gradients with high-frequency noise.

To confirm quantitatively this observation, we compute the Fourier signature of these gradients, quantifying the frequency power at each block. The resulting curves are shown in Fig. A.4. The gradients coming from a ResNet50 with MaxPooling (blue curves) contain all more power in the high-frequencies, compared to gradients issued from an architecture with AveragePooling (red curves). It's worth noting here that even the gradients after an AveragePooling contain some artifacts in the high-frequencies (illustrated by the tip of the curves going up). These artifacts may come from striding mechanisms and carried throughout layers.

We finally compare the contribution of downsampling operations in the frequency content of the gradients with all the other layers. Fig. A.5 represents the power/frequency slope of the gradient at each layer, from the prediction to the input. The blue curve represents a pre-trained ResNet50V2 with MaxPooling operation, while the black curve shows the slopes of an

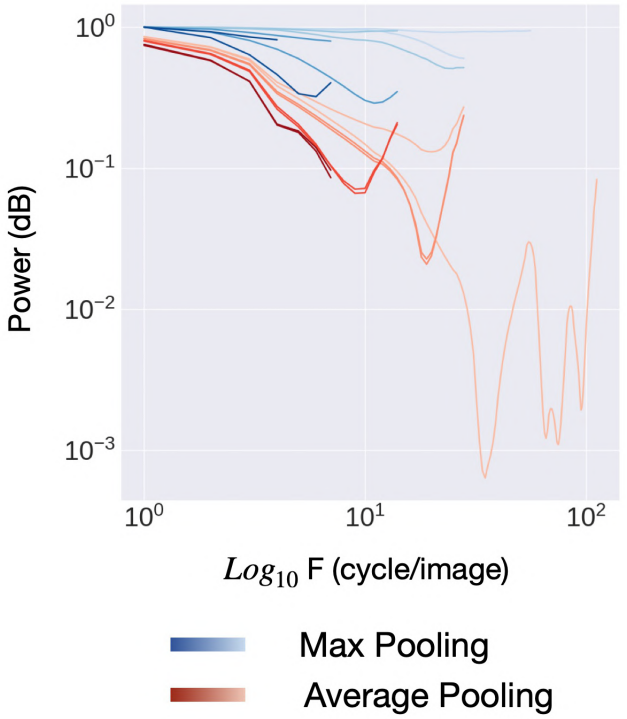

Figure A.4: **Fourier signature of the gradient after pooling operations in a ResNet50.** The gradients of a ResNet50 after MaxPooling operations (blueish curves) show a Fourier signature with more power in the high-frequencies compared to after an AveragePooling (reddish curves).

untrained network. These two curves show a comparable evolution per layer, confirming our hypothesis of an architectural mechanism, independent from the optimization. However, the red curve, representing the network with AveragePooling as a downsampling operation, shows a significant reduction of the slopes at each layer. We highlight here the downsampling steps – separation of each block – and invite the reader to compare the slope inside each block only. Indeed, the slopes between blocks are not comparable due to a different image size, and therefore a different length of the Fourier signature. The specific contribution of the pooling in Fig. A.5 corresponds to the first points of each block. The other layers inside each block do not appear to have a significant impact: the slopes remain stable inside each block.

Downsampling operations in a VGG16. Due to the large number of layers in a ResNet50V2, we perform the same analysis as above on a VGG16 to have a better understanding of the Fourier footprint of the downsampling operations. Fig. A.6 reproduces the results obtained with ResNet50V2 (Fig. A.5. Compared to the above, the operations other than the pooling seem to smooth the gradient (the slopes of the trained VGG16 – blue curve – decrease inside each block). However, we still observe a big gap at the beginning of each block due to the MaxPoolings.

We next plot in Fig. A.7 the Fourier signature of each layer of a VGG16 with MaxPooling (left panel) or AveragePooling (right panel). On both panels, the red curves represent the Fourier signature of the gradient after a pooling operation and the light blue curves correspond to convolutions. On the right panel, we observe a clear difference between the Fourier signature of the MaxPooling compared to the convolutions: the MaxPooling generates a Fourier signature with an increasing power in the higher frequencies. Conversely, the AveragePooling actively decreases the power in these specific high-frequencies, generating Fourier signatures with a significantly reduced frequency power compared to convolutions. The visualization of these gradients at each block, Fig. A.8 confirms our quantitative analysis.

Finally, we vary the stride value of the MaxPooling operations in a trained VGG16 with MaxPooling2D. In Fig A.9, we find that performing a MaxPooling with no stride (meaning using a value of 1) reduces the amount of high-frequency content introduced per layer, compared to using a stride value superior to 1. However, increasing the stride value from 2 to 4 does not change significantly the amount of high-frequency content.

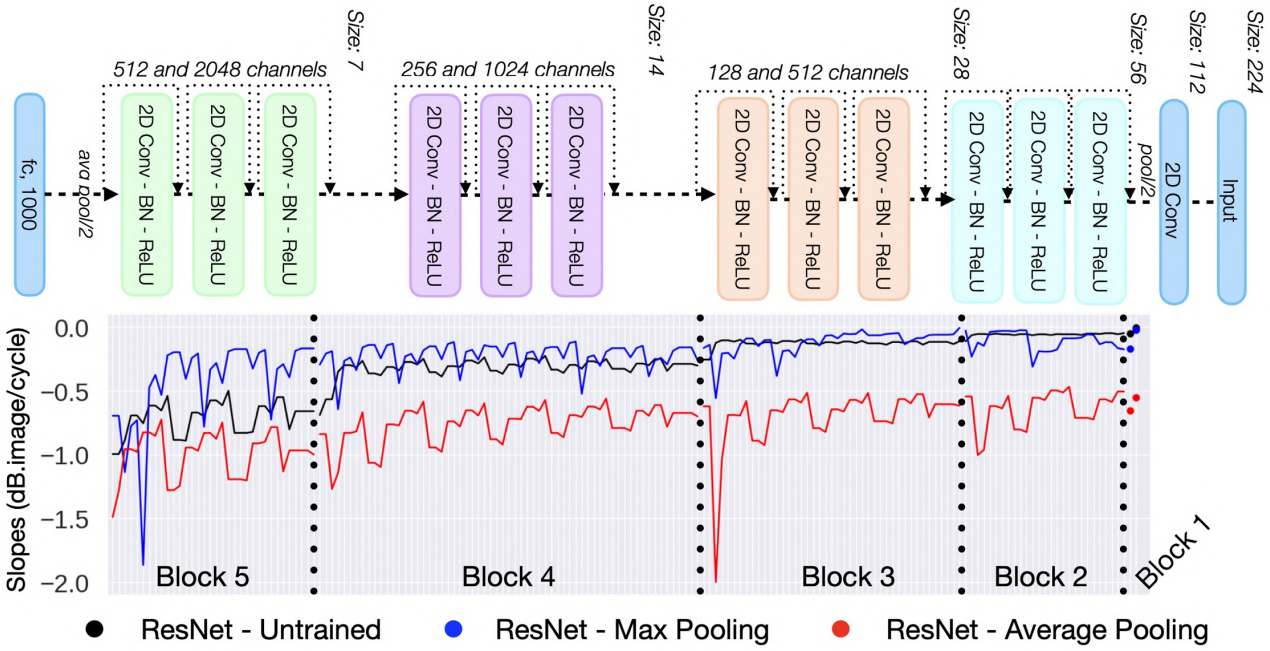

Figure A.5: **Frequency profiles in every layer of a ResNet50:** Evolution of the power/frequency slope of the gradients in each layer of a ResNet50 network. Here we consider a standard pre-trained ResNet50 (blue curve), an untrained ResNet50 network (black curve), and a ResNet50 in which we substitute MaxPooling with AveragePooling (red curve). The untrained network has the same frequency profile as the trained ResNet50. The ResNet50 with AveragePooling shows a lower power/frequency slope, indicating lower contamination of the high-frequency. All in all, this figure suggests that high frequency in the ResNet50 is not a by-product of the training but rather a consequence of the MaxPooling operation.

A.4. The importance of an adaptive sigma

In Fig. A.10, we show the evolution of the faithfulness w.r.t the value of σ. We select the optimal cutoff frequency $\sigma\star$ by maximizing the faithfulness. These values range from 224 to 2 (we plot the values here starting at 160 for the sake of visualization). We can observe that in most cases, the faithfulness score increases significantly after filtering the gradient. We also observe a pattern of evolution similar for the methods belonging to the same sub-categories (i.e. SmoothGrad, SquareGrad, VarGrad). Additionally, the filtering is more useful for convolutional networks (ResNet50V2, ConvNeXt) than for ViT. This can be explained by the absence of MaxPooling or striding in transformer architectures and smoother gradients. In Fig. A.11, we plot the fidelity metric using the same σ value as in Fig. A.10. The pattern of evolution is significantly different, and the σ value maximizing the faithfulness metric does not always correspond to the optimal σ value for the fidelity metric. However, it is possible to optimize σ on the fidelity metric as there is at least one value per model and method which is increased compared to the absence of filtering.

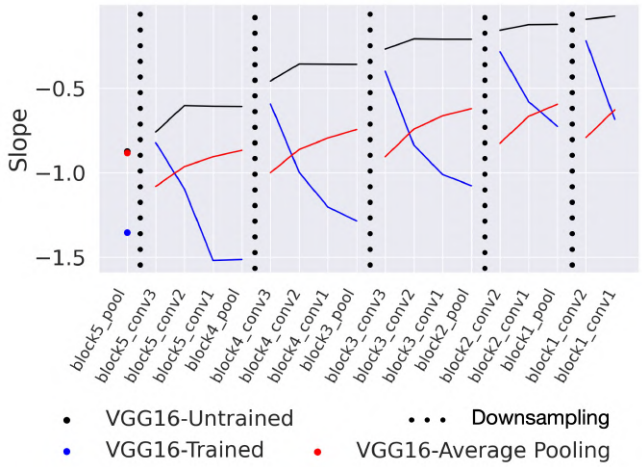

Figure A.6: **Frequency profiles in every layer of a VGG16:** Evolution of the power/frequency slope of the gradients in each layer of a VGG16 network. Here we consider a standard pre-trained VGG16 (blue curve), an untrained VGG16 network (black curve), and a VGG16 in which we substitute MaxPooling with AveragePooling (red curve). The untrained network has a power/frequency slope slightly higher than the pre-trained version. The VGG16 with AveragePooling shows a lower power/frequency slope at the beginning of each block, indicating lower contamination of the high-frequency. All in all, this figure suggests that high frequency in the VGG16 is not a by-product of the training but rather a consequence of the MaxPooling operation.

A.5. Qualitative examples

We provided in Fig. A.12 some qualitative examples of **FORGrad**. For several White-box methods, we show the produced attribution map before and after applying **FORGrad**. All the resulting explanations concord with the metrics, showing a clear improvement in quality, and smoother attribution maps.

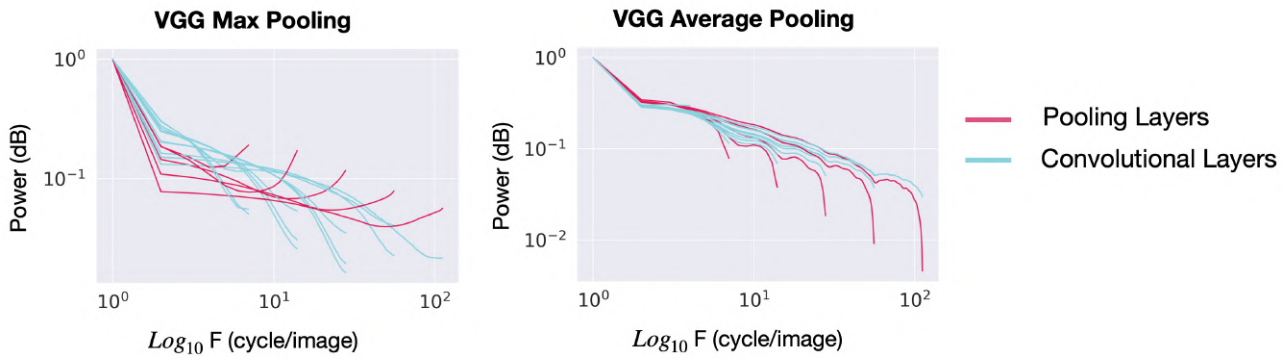

Figure A.7: **Fourier signature of the gradient after each layer in a VGG16.** The gradients of a VGG16 after MaxPooling operations (right panel – red curves) show a Fourier signature with more power in the high-frequencies compared to convolution layers in the whole architecture. Comparatively, replacing MaxPooling with AveragePooling lowers the power in the high-frequencies, resulting in a signature with lower high-frequency power than convolution operations.

A.6. Control experiments

We provide in this section some control experiments on **FORGrad**.

Controlling the metrics. As we use two main metrics to compute our quantitative results, we perform control experiments ensuring none of the metrics are biased towards attribution maps containing low or high-frequencies. To do so, we produce random explanations of very low frequencies by drawing a 2D Gaussian at a random position in an empty map. We generate the high-frequency equivalent by dispatching all the active pixels of the low-frequency map in another empty map. Some examples are shown in Fig. A.13. We then compute the scores given by each metric for 1000 maps of each category. The results are reported in Tab 3. We add the scores obtained with Saliency as a baseline. The results show a potential bias towards high-frequency explanations explaining why the scores obtained with **FORGrad** are a bit less salient. However, no bias appears with faithfulness, confirming the real improvement of **FORGrad** on each of the white-box methods.

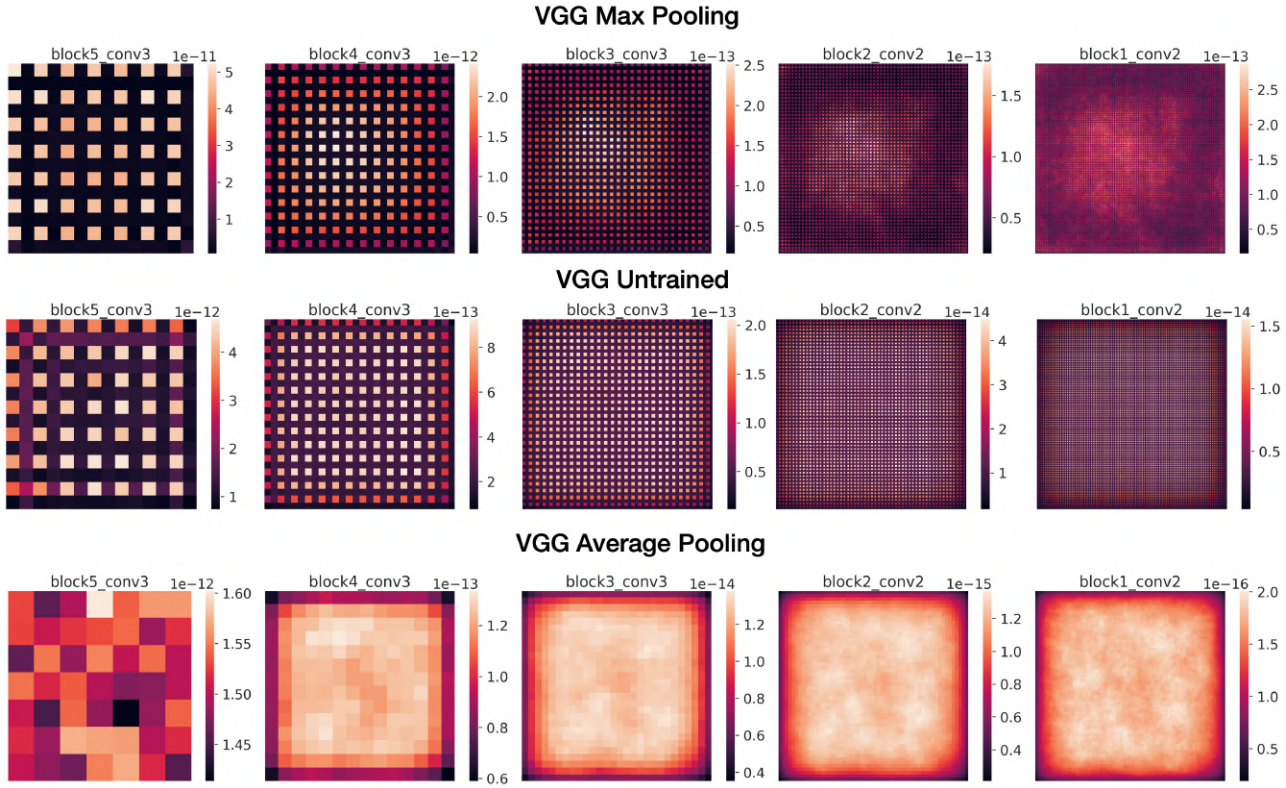

Figure A.8: Visualization of the gradient (w.r.t. the input) probed at different layers. We observe that both in the VGG MaxPooling and the VGG untrained, a checkerboard signal emerges in the gradients. Such a signal might cause an increased power in high frequencies. In contrast, there is no such signal in the VGG Average Pooling. It suggests that MaxPooling operations, creating such a checkerboard pattern, might contaminate the gradients with high-frequency artifacts.

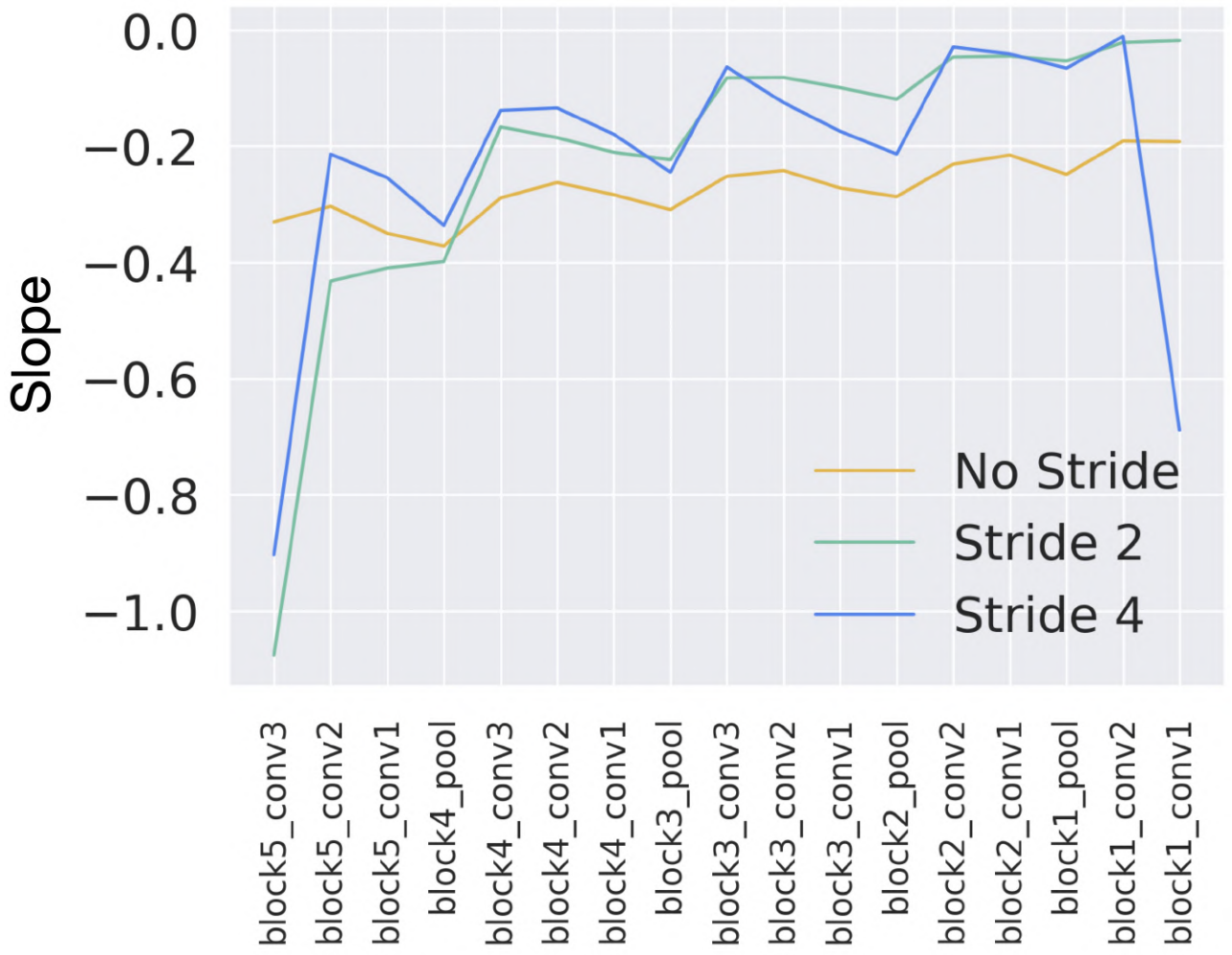

Figure A.9: Effect of the stride value of MaxPooling in the VGG16. We observe that performing a MaxPooling with a stride of 1 (akin to no stride), reduces the amount of high-frequency artifacts included in the gradients of the model. Varying the stride value (2 or 4) does not change the high-frequency power.

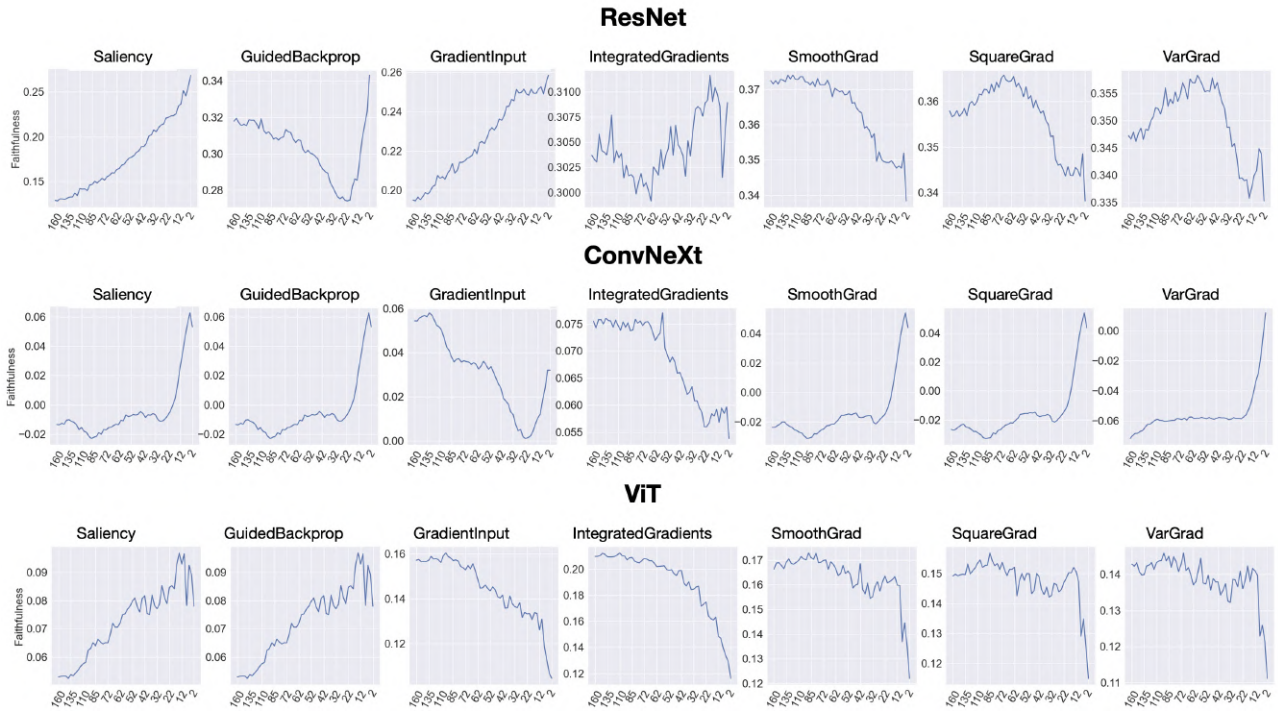

Figure A.10: **Evolution of the faithfulness depending on the cut-off frequency σ.** The faithfulness evolves differently per methods and models depending on the cut-off frequency σ. The filtering tends to be more beneficial for convolutional networks (ResNet50 and ConvNeXt) compared to ViT due to the absence of pooling on the latter family of models. Consistently, filtering improves the faithfulness score or at least does not reduce it.

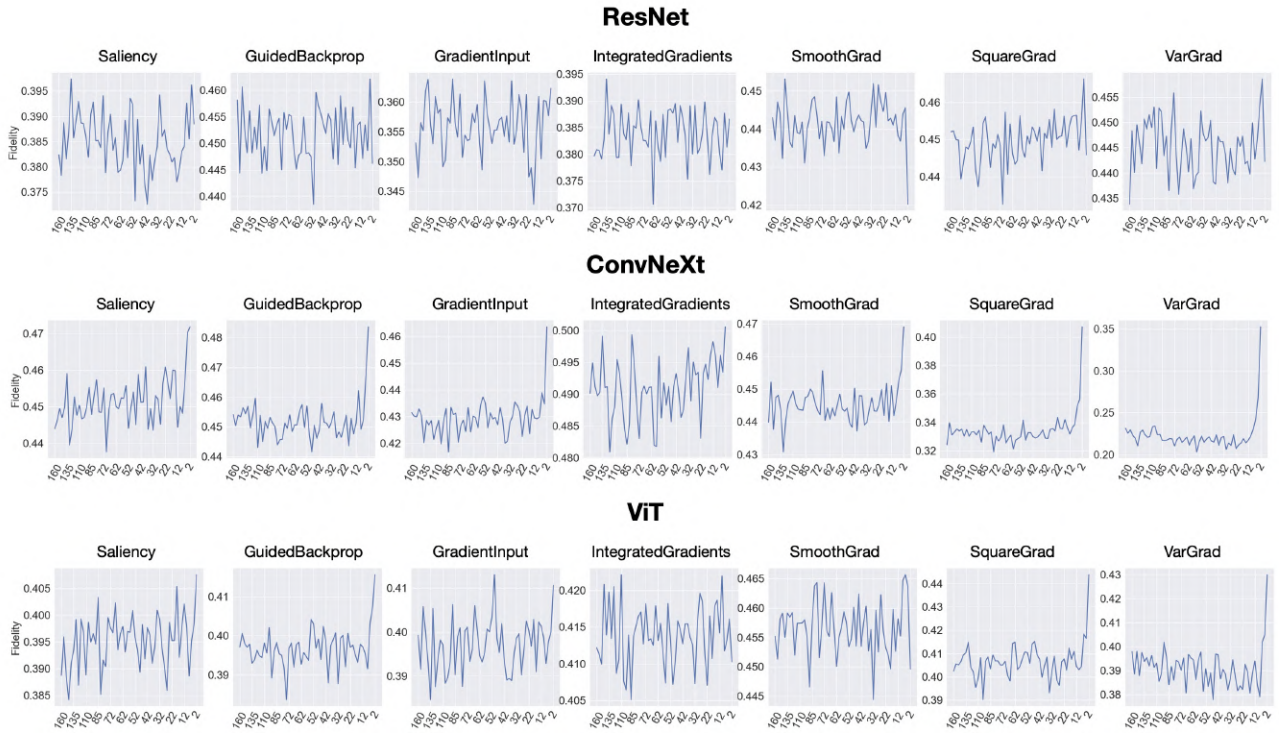

Figure A.11: **Evolution of the fidelity depending on the cut-off frequency σ.** The faithfulness evolves differently per methods and models depending on the cut-off frequency σ. Consistently, filtering improves the faithfulness score or at least does not reduce it.

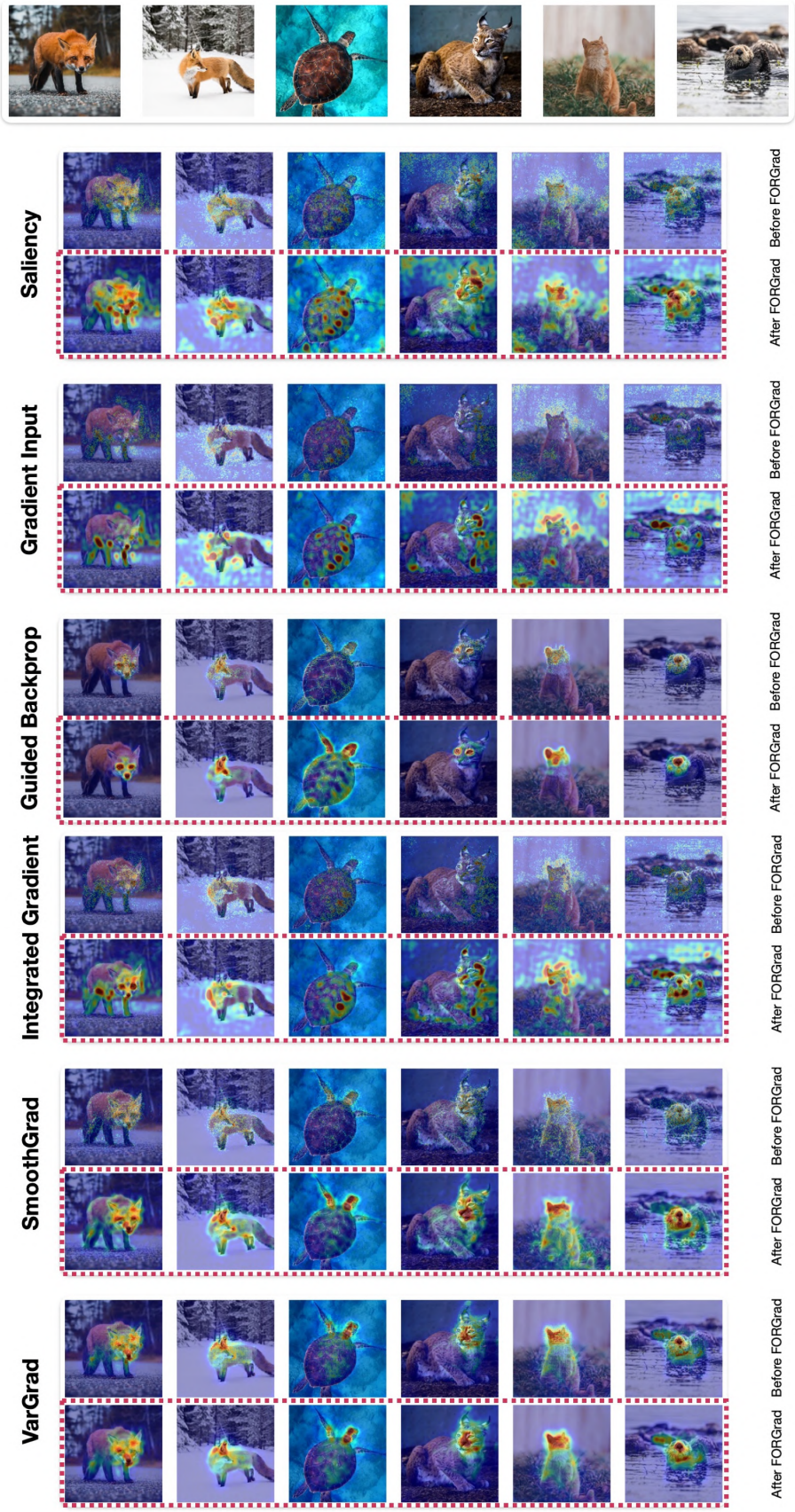

Figure A.12: **Qualitative examples of several white-box methods before and after FORGrad**

Figure A.13: **Visualization of low and high-frequency random attribution maps.** The low-frequency random maps consist of 2D Gaussian randomly located in an image (the background images are for the sake of visualization). The high-frequency equivalent consists of dispatching the pixels of the Gaussian anywhere on the image.

Method name	Deletion	Insertion	Faithfulness	μ Fidelity
Saliency	0.127	0.316	0.189	0.390
Random low-freq.	0.205	0.187	-0.018	0.140
Random high-freq.	0.138	0.179	0.041	0.310

Table 3: **Faithfulness and μFidelity scores on random explanation of high or low frequencies.**We generate random explanation containing either high or low-frequency information and compute their scores given by each of our two metrics. The faithfulness metric shows no apparent bias while μFidelity tends to favor high-frequency explanations.

Sanity check for Saliency. The second experiment reproduces the sanity check proposed by Adebayo et al. (2018). This experiment consists of progressively randomizing the weights of the network and recomputing the attribution map. As a result, the methods that are altered by the randomization pass the test, while the methods producing the same result before or after randomization fail. In Fig. A.14, we show that **FORGrad** does not change the results of this sanity check. In other words, the methods that failed before still do, and conversely, the ones that passed the test are still affected by the randomization.

Statistical effect of FORGrad. To show the statistical effect of **FORGrad** for each metric we performed a Bayesian ANOVA (see details below). We analyzed the metrics by means of Bayesian ANOVAs, considering BEFORE-AFTER and ATTRIBUTION-METHODS as fixed factors, and MODELS as a random term. In all analyses, we computed Bayes Factors (BF) as the ratio between the models testing the alternative against the null hypothesis. The alternative hypothesis states that the factor is important for explaining the variance in the data (i.e., there is a significant difference between conditions, for example before and after). All BFs are denoted as BF10. In practice, BFs provide substantial (BF¿ 5) or strong (BF¿ 10) evidence in favor of the alternative hypothesis, and low BF (BF¡ 0.5) suggests a lack of effect (Masson, 2011). Confirming our results, we found a strong difference in the BEFORE-AFTER factor (comparing scores before and after FORGrad) in the INS metric (BEFORE-AFTER factor BF10=11.33, error= 1%), and moderate evidence for the DEL metric (BEFORE-AFTER factor BF10=1.79, error= 2%). All in all, considering the small number of samples used for this analysis, we are very confident that the results of the statistical analysis confirm a significant difference between the scores obtained before and after applying FORGrad on the different attribution methods.

Standard Deviation. We additionally include in Table 7 the standard deviation computed on the scores obtained on each image, for each attribution and metric on ResNet50.

A.7. Quantitative results of FORGrad

Illustrating the contribution of FORGrad. We summarize here our contribution with **FORGrad**. We notice a difference in the Fourier spectra of the different categories of attribution methods. We demonstrate that this difference – white-box methods show more power in the high-frequencies – is the reason why black and gray-box methods outperform the first category of methods. The observation is illustrated in the following figure (Fig. A.15): white-box methods are mainly located at the right bottom of the plots. **FORGrad** "repairs" these methods by filtering the high-frequency artifacts in the gradient, bringing them to the upper-left corner of the plots.

Quantitative evaluation of attribution methods on ResNet50, VIT and ConvNext. We additionally report here the faithfulness and fidelity scores of white-box methods before and after **FORGrad** and compare them with gray and black-box methods for ResNet50V2, ConvNeXt, and ViT. GradCAM methods can't be tested on ViT because they are based on convolution so are limited to CNNs. For the 3 models, before **FORGrad** the best methods are always black-box. Even though **FORGrad** does not allow to systematically beat them, it brings white-box methods back to competitive performance. Considering their computational efficiency, **FORGrad** rends white-box methods very attractive given the accuracy/efficiency tradeoff.

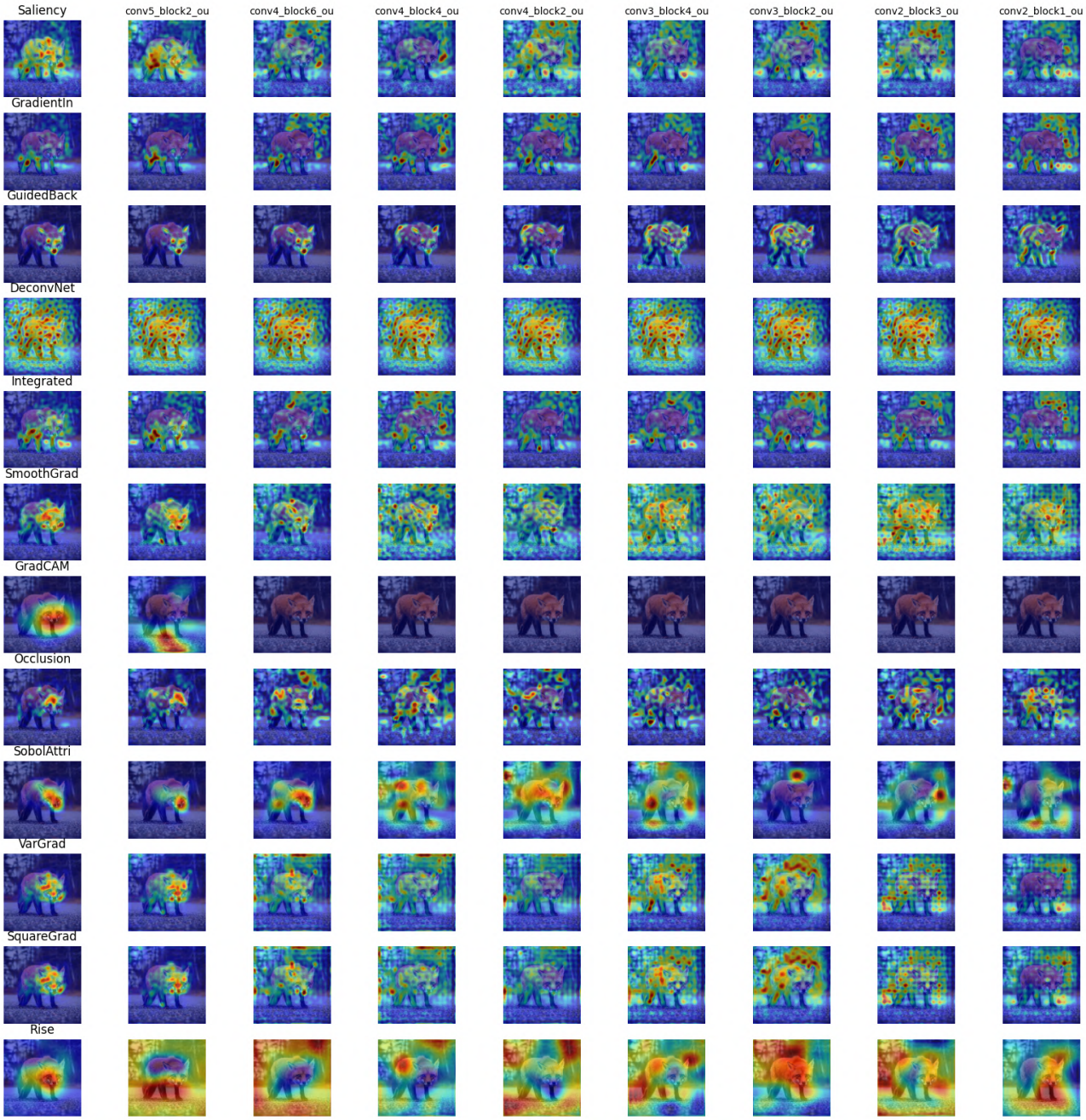

Figure A.14: **Sanity check of the methods after FORGrad.** We perform the sanity check for attribution methods as presented in (Adebayo et al., 2018) on a ResNet50. The first column corresponds to the original heatmap generated by each method. The other columns show the heatmap produced after randomizing the weights at various layers of the model. Methods that consistently offer a comparable explanation to the original one are deemed to fail the test.

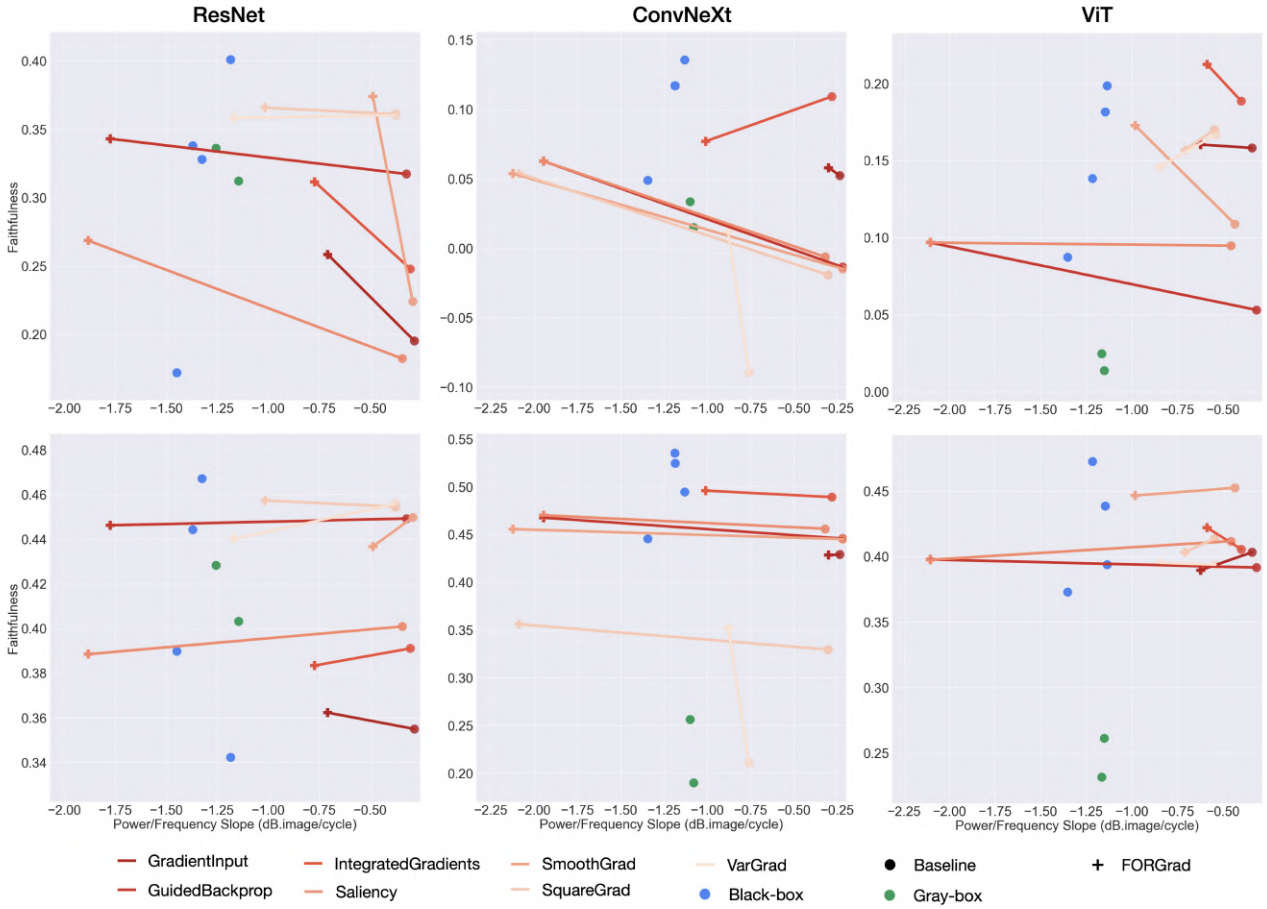

Figure A.15: **Effect of FORGrad on white-box attribution methods.** Before applying **FORGrad** most white-box methods (reddish points) contain a lot of power in the high-frequencies (x-axis) but a lower explainability score (y-axis) than other types of methods (blue and green points). **FORGrad** corrects white-box methods by low-pass filtering the gradient (reducing the high-frequency power), bringing them back to a faithfulness closer to gray and black-box methods.

		ResNet50				ConvNeXT				ViT			
		Faith.(↑)	μFid.(↑)	Stab.(↓)	Time(↓)	Faith.(↑)	μFid.(↑)	Stab.(↓)	Time(↓)	Faith.(↑)	μFid.(↑)	Stab.(↓)	Time(↓)
Gradient-based	Saliency(Simonyan et al., 2013)	0.18	0.40	0.67	0.78	-0.01	0.46	7e-05	<u>1.89</u>	0.09	0.41	0.06	**0.45**
	Saliency*	0.27	0.39	0.53	0.89	0.06	0.47	3e-05	**1.64**	0.10	0.40	0.03	<u>0.46</u>
	Guidedbackprop(Ancona et al., 2018)	0.31	0.45	0.28	8.25	-0.02	0.45	4e-05	19.3	0.05	0.39	0.04	5.48
	Guidedbackprop$\star$	0.34	0.45	0.22	7.05	0.06	0.47	3e-05	17.8	0.10	0.40	0.03	6.41
	GradInput(Ancona et al., 2018)	0.20	0.36	0.42	**0.73**	0.05	0.43	0.005	1.93	0.16	0.40	0.02	0.53
	GradInput$\star$	0.26	0.36	0.35	<u>0.77</u>	0.06	0.43	0.001	1.93	0.16	0.39	0.006	0.59
	Int.Grad(Sundararajan et al., 2017)	0.24	0.39	0.72	42.7	0.11	0.49	0.16	93.5	0.19	0.41	0.12	24.6
	Int.Grad$\star$	0.31	0.38	0.76	41.8	0.07	0.50	0.14	93.2	**0.21**	0.42	0.12	23.4
	SmoothGrad(Smilkov et al., 2017)	0.23	0.45	0.22	46.6	0.02	0.45	4e-05	104	0.11	<u>0.45</u>	0.03	26.4
	SmoothGrad$\star$	<u>0.37</u>	0.44	0.21	48.3	0.05	0.46	3e-05	103	0.17	<u>0.45</u>	0.05	26.9
	VarGrad(Seo et al., 2018)	0.36	<u>0.46</u>	**0.003**	41.5	-0.09	0.21	1e-10	93.7	0.16	0.39	0.001	27.1
	VarGrad$\star$	0.35	0.44	<u>0.004</u>	40.6	0.01	0.35	4e-13	90.6	0.15	0.40	0.0003	23.4
	SquareGrad(Seo et al., 2018)	0.36	0.45	**0.003**	42.1	-0.02	0.33	2e-09	93.4	0.17	0.41	0.001	20.5
	SquareGrad$\star$	0.36	<u>0.46</u>	0.005	40.9	0.05	0.36	4e-10	90.7	0.17	0.41	0.001	19.3
Prediction-based	GradCAM(Selvaraju et al., 2017a)	0.31	0.40	0.31	5.24	0.03	0.26	0.005	6.80	n.a	n.a	n.a	n.a
	GradCAM++(Chattopadhay et al., 2018)	0.33	0.43	0.34	4.61	0.02	0.19	0.001	7.06	n.a	n.a	n.a	n.a
	Occlusion(Ancona et al., 2018)	0.17	0.39	0.60	368	0.05	0.45	0.05	782	0.09	0.37	0.09	183
	HSIC(Novello et al., 2022)	0.33	**0.47**	0.45	456	0.08	**0.54**	0.14	1358	<u>0.20</u>	**0.47**	0.14	248
	Sobol(Fel et al., 2021b)	0.34	**0.47**	0.47	578	<u>0.12</u>	<u>0.53</u>	0.16	1156	0.14	**0.47**	0.05	342
	RISE(Petsiuk et al., 2018)	**0.41**	0.34	0.55	626	**0.13**	0.49	0.15	1534	0.18	0.44	0.08	487

Table 4: **Results on Faithfulness metrics**. Faithfulness and Fidelity scores were obtained on 1,280 ImageNet validation set images, on an Nvidia T4 (For Faithfulness and Fidelity, higher is better). Time in seconds corresponds to the generation of 100 (ImageNet) explanations on an Nvidia T4. The first and second best results are in **bold** and <u>underlined</u>.

	ResNet50		ConvNeXT		ViT	
	Original	**FORGrad**	Original	**FORGrad**	Original	**FORGrad**
1	VarGrad (0.82)	**SquareGrad⋆ (0.82)**	Sobol (0.49)	**Saliency⋆ (0.53)**	SquareGrad (0.579)	**SquareGrad⋆ (0.579)**
2	SquareGrad (0.81)	**VarGrad⋆ (0.82)**	HSIC (0.48)	**GuidedBackprop⋆ (0.53)**	Sobol (0.56)	**SmoothGrad⋆ (0.57)**
3	GuidedBackprop (0.48)	**SmoothGrad (0.60)**	GradInput (0.475)	**SmoothGrad⋆ (0.51)**	VarGrad (0.549)	Sobol (0.56)
4	SmoothGrad (0.46)	**GuidedBackprop⋆ (0.58)**	Rise (0.47)	Sobol (0.49)	GradInput (0.54)	**VarGrad⋆ (0.55)**
5	GradCAM++ (0.42)	GradCAM (0.42)	SmoothGrad (0.47)	**GradInput⋆ (0.489)**	Rise (0.54)	GradInput (0.544)

Table 5: **Global ranking before (original) and after FORGrad.** For each model, we show the 5 attribution methods with the highest metrics, before and after applying **FORGrad**. The explanation maps were computed on 1280 images from the validation set of ImageNet, based on an aggregation of the three metrics computed by $F(\boldsymbol{x}, \boldsymbol{\varphi}) + \mu F(\boldsymbol{x}, \boldsymbol{\varphi}) - S(\boldsymbol{x}, \boldsymbol{\varphi})$.

Additional quantitative control measurements. We finally include additional experiments to control the robustness of **FORGrad** on the different metrics and parameters. Table 6 summarizes the results of **FORGrad** when the baseline value of Insertion and Deletion metrics is changed from 0 to uniform noise between -1 and 1. We can see that the scores indeed differ, but the effectiveness of the method does not. Table 7 evaluates the effect of applying **FORGrad** (i.e. low-pass filtering using an optimal cut-off value) after computing the attribution map instead of filtering the gradient prior to generating the attribution map († symbol) on ResNet50. Even though the metrics improve compared to the absence of filtering, filtering the gradient before computing the attribution map is the most optimal.

		ResNet50		
		Del.(↓)	Ins.(↑)	Faith.(↑)
White-box	Saliency(Simonyan et al., 2013)	0.08	0.27	0.17
	Saliency⋆	0.09	**0.37**	**0.28**
	Guidedbackprop(Ancona et al., 2018)	0.05	0.33	0.28
	Guidedbackprop⋆	0.07	**0.39**	**0.32**
	GradInput(Shrikumar et al., 2017)	0.08	0.23	0.15
	GradInput⋆	0.11	**0.33**	**0.35**
	Int.Grad(Sundararajan et al., 2017)	0.11	0.17	0.06
	Int.Grad⋆	**0.10**	0.17	**0.07**
	SmoothGrad(Smilkov et al., 2017)	0.13	0.31	0.18
	SmoothGrad⋆	0.13	**0.34**	**0.21**
	VarGrad (Adebayo et al., 2018)	0.12	0.28	0.16
	VarGrad⋆	0.13	**0.33**	**0.20**
	SquareGrad(Seo et al., 2018)	0.12	0.29	0.17
	SquareGrad⋆	0.13	**0.33**	**0.20**
Black & Gray-box	GradCAM(Selvaraju et al., 2017a)	0.10	0.40	0.30
	GradCAM++(Chattopadhay et al., 2018)	0.08	0.41	0.33
	Occlusion(Ancona et al., 2018)	0.08	0.28	0.20
	HSIC(Novello et al., 2022)	0.06	0.42	0.36
	Sobol(Fel et al., 2021b)	0.07	0.40	0.33
	RISE(Petsiuk et al., 2018)	0.07	0.41	0.34

Table 6: **Results on Deletion, Insertion, and Faithfulness metrics**. Deletion, Insertion, and Faithfulness scores were obtained on 1,280 ImageNet validation set images, on an Nvidia T4 (For Insertion and Faithfulness, higher is better). The scores that are improved by **FORGrad** are in **bold**

		ResNet50			
		Faith.(↑)	μFid.(↑)	Stab.(↓)	Time(↓)
White-box	Saliency(Simonyan et al., 2013)	0.18 (±0.03)	0.40 (±0.06)	0.67 (±0.3)	0.78
	Saliency†	0.28 (±0.03)	0.41 (±0.06)	0.56 (±0.23)	0.87
	Saliency⋆	0.28 (±0.02)	0.39 (±0.06)	0.53 (±0.16)	0.89
	Guidedbackprop(Ancona et al., 2018)	0.31 (±0.07)	0.45 (±0.06)	0.28 (±0.11)	8.25
	Guidedbackprop†	0.33 (±0.03)	0.43 (±0.07)	0.29 (±0.06)	7.94
	Guidedbackprop⋆	0.35 (±0.03)	0.45 (±0.07)	0.22 (±0.06)	7.05
	GradInput(Shrikumar et al., 2017)	0.2 (±0.04)	0.36 (±0.06)	0.42 (±0.2)	**0.73**
	GradInput†	0.22 (±0.03)	0.37 (±0.07)	0.39 (±0.12)	0.75
	GradInput⋆	0.26 (±0.01)	0.36 (±0.06)	0.35 (±0.11)	<u>0.77</u>
	Int.Grad(Sundararajan et al., 2017)	0.24 (±0.05)	0.39 (±0.06)	0.72 (±0.21)	42.7
	Int.Grad†	0.26 (±0.03)	0.40 (±0.06)	0.72 (±0.20)	41.3
	Int.Grad⋆	0.31 (±0.03)	0.38 (±0.06)	0.76 (±0.20)	41.3
	SmoothGrad(Smilkov et al., 2017)	0.23 (±0.04)	0.45 (±0.07)	0.22 (±0.06)	46.6
	SmoothGrad†	0.32 (±0.02)	0.43 (±0.06)	0.22 (±0.08)	47.0
	SmoothGrad⋆	<u>0.37</u> (±0.02)	0.44 (±0.06)	0.21 (±0.04)	48.3
	VarGrad (Adebayo et al., 2018)	0.36 (±0.05)	<u>0.46</u> (±0.07)	**0.003** (±0.001)	41.5
	VarGrad†	0.34 (±0.03)	0.44 (±0.06)	0.009 (±0.003)	41.1
	VarGrad⋆	0.35 (±0.03)	0.44 (±0.06)	<u>0.004</u> (±0.002)	40.6
	SquareGrad(Seo et al., 2018)	0.36 (±0.05)	0.45 (±0.07)	**0.003** (±0.001)	42.1
	SquareGrad†	0.36 (±0.03)	0.45 (±0.06)	0.01 (±0.003)	41.9
	SquareGrad⋆	0.36 (±0.03)	<u>0.46</u> (±0.06)	0.005 (±0.003)	40.9
Black & Gray-box	GradCAM(Selvaraju et al., 2017a)	0.31 (±0.04)	0.40 (±0.06)	0.31 (±0.13)	5.24
	GradCAM++(Chattopadhay et al., 2018)	0.33 (±0.05)	0.43 (±0.06)	0.34 (±0.14)	4.61
	Occlusion(Ancona et al., 2018)	0.20 (±0.03)	0.39 (±0.06)	0.6 (±0.25)	368
	HSIC(Novello et al., 2022)	0.33 (±0.04)	**0.47** (±0.07)	0.45 (±0.19)	456
	Sobol(Fel et al., 2021b)	0.34 (±0.03)	**0.47** (±0.05)	0.47 (±0.21)	578
	RISE(Petsiuk et al., 2018)	**0.41** (±0.05)	0.34 (±0.06)	0.55 (±0.17)	626

Table 7: **Results on Faithfulness metrics**. Faithfulness, Fidelity, and Stability scores were obtained on 1,280 ImageNet validation set images, on an Nvidia T4 (For Faithfulness and Fidelity, higher is better). Time in seconds corresponds to the generation of 100 (ImageNet) explanations on an Nvidia T4. The first and second best results are in **bold** and <u>underlined</u>. † = **FORGrad** computed on the attribution map. ⋆ = **FORGrad** computed on the gradients.