On the Complexity-Faithfulness Trade-off of Gradient-Based Explanations

Amir Mehrpanah Matteo Gamba Kevin Smith Hossein Azizpour
KTH Royal Institute of Technology, Sweden
{amirme, mgamba, ksmith, azizpour}@kth.se

Abstract

ReLU networks, while prevalent for visual data, have sharp transitions, sometimes relying on individual pixels for predictions, making vanilla gradient-based explanations noisy and difficult to interpret. Existing methods, such as Grad-CAM, smooth these explanations by producing surrogate models at the cost of faithfulness. We introduce a unifying spectral framework to systematically analyze and quantify smoothness, faithfulness, and their trade-off in explanations. Using this framework, we quantify and regularize the contribution of ReLU networks to high-frequency information, providing a principled approach to identifying this trade-off. Our analysis characterizes how surrogate-based smoothing distorts explanations, leading to an "explanation gap" that we formally define and measure for different post-hoc methods. Finally, we validate our theoretical findings across different design choices, datasets, and ablations.[1]

1. Introduction

The growing complexity of deep networks necessitates explainability, especially in safety-critical applications. Gradient-based explanation methods are widely used in computer vision for this aim. These methods, however, suffer from two main shortcomings: lack of *smoothness* or *faithfulness*, the trade-off of which is the focus of our study.

Explanation complexity. A crucial property for explanations to be human-understandable is their simplicity, which translates to smoothness[2] in saliency-based explanations. However, both theoretical [19, 28] and empirical [47, 54] studies suggest that ReLU networks, the most common variant of deep networks used for visual data, have sharp transitions, sometimes relying on individual pixels for predictions [57]. This sharpness can result in complex gradient-based explanations that hinder interpretability. Thus, as our preliminary step, we establish a formal connection between a network's architecture and complexity of its explanations. With this formalism, we particularly study ReLU networks.

Explanation faithfulness. To reduce the explanation complexity, post-hoc techniques produce smooth surrogates during the explanation phase [41], at the cost of lower faithfulness to the original network, leading to an inherent trade-off between explanation complexity and faithfulness. Therefore, it is vital to carefully consider this trade-off when developing, choosing, or using post-hoc methods [1, 23]. While several metrics are proposed to measure complexity (*e.g.*, entropy-based methods [6]) and faithfulness (*e.g.*, pixel removal scores [6]), these metrics disjointly measure one of the two aspects and are influenced by extraneous factors such as the choice of baseline and removal order [33, 60]. These caveats of existing metrics impede a principled analysis of the complexity-faithfulness trade-off [14, 46]. Our goal is to formally study this trade-off. Specifically, we introduce consistently-defined measures of complexity and faithfulness with which we empirically analyze the existing post-hoc explanation methods, and our method of mitigation of sharpness in ReLU networks.

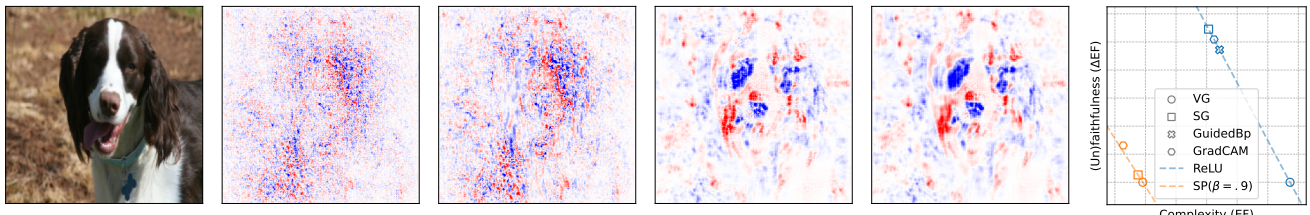
Our Approach. We adopt a *spectral* perspective to formally establish a relationship between a network's architecture and its gradient-based explanation. We first measure explanation complexity via spatial high-frequency content of input gradients (Sec. 3.2), and formally link it to the high-frequency components, or sharp changes, of a network (Sec. 3.3). This connection, importantly, allows for characterization and improvement of network architectures for their explanation complexity, which we showcase in Sec. 3.4. Finally, in Sec. 4, we leverage our formalism in a unified framework to analyze the complexity-faithfulness trade-off, *i.e.*, by measuring complexity as "expected frequency" and faithfulness as the "gap" in the expected frequency of the original and surrogate functions. We use this framework to analyze (smoothed versions of) ReLU networks and various post-hoc explanations; see Fig. 1.

Our Contributions. We summarize in the following list:

- **A Framework for Formal Analysis:** We introduce a framework based on a formal connection between a network's power spectrum tail and the complexity of its ex-

[1] https://github.com/Amir-Mehrpanah/On-the-Complexity-Faithfulness-Trade-off-of-Gradient-Based-Explanations-ICCV25/

[2] We may refer to smoothness and complexity interchangeably.



(a) Input (b) ReLU:VG (c) ReLU:SG (d) SP($\beta = .9$):VG (e) SP($\beta = .9$):SG (f)

Figure 1. **Exploring the Faithfulness-Complexity Trade-Off.** This figure shows the trade-off typically made between faithfulness and complexity in post-hoc explanation methods. **(b)** presents VanillaGrad (VG) explanations for a ReLU network, which appear grainy due to the network's reliance on high-frequency information. To mitigate this, post-hoc methods like SmoothGrad (SG), shown in **(c)**, remove high-frequency components by creating surrogates, thereby introducing a trade-off between complexity and faithfulness, quantified by EF and Δ EF, respectively. This trade-off for a ReLU network is represented by the dashed blue line in **(f)**. In this work, we propose a framework which can control a ReLU network's reliance on high-frequency information by modifying the tail of its power spectrum. This is achieved by training a network with a smooth parameterization (SP) of ReLU, obtained by convolving ReLU with a Gaussian function, where the standard ReLU corresponds to $\beta \to \infty$. **(d)** and **(e)** show the VG and SG explanations for such a network, denoted as SP($\beta = 0.9$). As shown, reducing the network's dependence on high-frequency components leads to VG explanations with lower complexity, enabling alternative trade-offs. One such trade-off is depicted by the dashed orange line in **(f)**—see Fig. 14 for more examples.

planations, which provides a foundation for analyzing the complexity-faithfulness trade-off. Using this framework, we analyze the role of ReLU in inducing sharp transitions in the network, leading to complex explanations.

- **The Expected Frequency (EF):** We introduce a metric to *quantify* a network's reliance on high-frequency information. We reduce EF by convolving ReLU with a Gaussian function, hence *controlling* the complexity of VanillaGrad explanations (Sec. 3).
- **The Explanation Gap:** We introduce the *"explanation gap"*, a measure for quantifying the discrepancy between explanations produced for surrogates and the original model, highlighting potential faithfulness risks associated with post-hoc explanation methods, in Sec. 4.
- **Empirical Validation:** We empirically validate our theoretical contributions across various datasets of varying input sizes, such as ImageNet, Imagenette, CIFAR10, and Fashion-MNIST. Moreover, we conduct several ablation studies, in Appendix B, including the effect of other design decisions such as skip connections and batch norm.

2. Related Works

Although inherent and post-hoc explainability are distinct concepts [48], both can rely on gradients as their explainer, as seen in methods like B-cos [11] and Integrated Gradients [58], respectively. We focus on methods that utilize gradients as their explainer, without making a distinction between post-hoc or inherent explainability.

2.1. Explanation for Smooth Surrogates

Our work is related to research on explainability methods that incorporate smooth surrogates, either as part of their design or in post-processing. Whether explicit or implicit, a common part in these methods is the use of smooth surrogates to improve visual quality.

One approach to creating smooth surrogates is through a *perturbation distribution*, which can involve noise in the input space [54], noise in the parameter space [10], input space traversal [31, 49], spatial Gaussian blur [11], or removing negative gradient values [32, 55].

Another method for creating surrogates is by using *subnetworks*, as in GradCAM [50], or by making *modifications* to the original structure, as seen in LRP [3], DeepLIFT [51], and their variants.

Throughout our work, we base our formal analysis on the simplest gradient-based method, VanillaGrad [53], which does not involve any form of smooth surrogate.

2.2. Spectral View of Explanation

Closely related to ours is the research that adopts a spectral perspective on explanations [35, 36, 59], where the Fourier basis is defined along the spatial dimensions, such as width and height. While these methods typically use optimization to impose certain spectral properties on the explanations, our work leverages the spectral view solely for measurements in spatial domain. Also, our work is connected to spectral view to networks' learnt functions [29, 40, 45].

2.3. Metrics for Evaluating Explanations

Evaluating explanations remains challenging, with quite a few quantitative techniques available. Prior work has assessed explanations through human studies [38] and auxiliary tasks to verify attribution relevance [64], while others test whether explanation features are meaningful to the model [13]. A common approach, called pixel removal score [6], generates out-of-distribution samples, prompting a computationally heavy remove-and-retrain strategy [26].

Moreover, the correlation between output and pixel removal score is used as faithfulness metric [6]. However, despite a high correlation, explanations for surrogate and original models can differ in high-frequency components [59]. Unlike these metrics, our formalism provides hyperparameter-free metrics with clearer intuition that better capture feature interactions.

3. Model Structure: A Root Cause for Explanation Complexity

As widely discussed in the literature, there is a recognized need to refine raw gradient-based explanations, commonly known as VanillaGrad, which are sometimes referred to as "noisy" [10, 54]. This stems from the human preference for simpler and more interpretable explanations.

Input gradients of deep networks often appear scattered and noisy Fig. 1b. However, an important question arises: *Are we truly observing noise in VanillaGrad explanations, or do these high-frequency variations stem from an underlying structural property of the network?*

In the literature, this noisiness is commonly referred to as "Explanation Complexity". In this section, we introduce a formalism to characterize a root cause for gradient-based explanations' complexity: *the network's architecture, which influences its explanation complexity*.

After introducing our notation in Sec. 3.2, we measure explanation complexity by a summary statistic of (high) spatial frequency content of input gradients. This is followed by establishing a formal connection between the high spatial frequency of input gradients and high frequency contents (or sharpness) of a network's function in Sec. 3.3. This central connection provides interesting implications for understanding and reducing explanation complexity.

In Sec. 3.4, we reveal one such important implication for the most common family of deep networks in the visual domain, *i.e*. ReLU networks. Finally, in Sec. 4, we leverage our formalism for a principled analysis of the explanation complexity and its trade-off with faithfulness to the model– two axes to evaluate explanation methods that are measured in a unified framework, for the first time, in this work.

3.1. Notations

To maintain generality and simplicity, we consider a scalar multivariate classifier denoted by $f : \mathbb{R}^n \to \mathbb{R}$. In the context of gradient-based explainability, we are particularly interested in computing the gradient of the network f trained on a dataset $\mathcal{X}$ with respect to an input sample x, *i.e*. $x \mapsto \nabla_x f$. For brevity, we may omit the x-subscript on the gradient operator, as gradients are consistently taken with respect to the input throughout this article.

Building on prior work that adopts a spectral perspective on neural networks [45], we introduce a compact notation, to represent the Fourier basis. Under general assumptions, such as absolute integrability ($\int |f(x)|dx < \infty$), a function $f(x)$ can be decomposed into its frequency components: $\widehat{f}(\omega) = \int f(x)\psi(\omega, x)dx$, yielding an alternative representation of the function in the frequency domain, which we denote by $\widehat{f}(\omega)$. To analyze the frequency content of the explanations, we will also use the power spectrum (PS) of the signals, defined as the squared magnitude of the Fourier transform, $\mathrm{S}_f(\omega) = |\widehat{f}(\omega)|^2$.

A subtle technical detail arises in defining the Fourier basis along the spatial dimensions (orthogonal to the gradient direction) and along the pixel values (parallel to the gradient direction). We will focus only on the tail behavior of two power spectra in our analysis: 1) the Tail of Spatial Power Spectrum *(TSPS)* of input-gradient of a network, where spatial means that the Fourier basis are defined along the spatial dimensions, *i.e*. width and height of the input-gradient. 2) the Tail of Power Spectrum *(TPS)* of a network, where the Fourier bases are defined along the pixel values.

3.2. Explanation Complexity and TSPS of Gradient

The noisier the explanation, the heavier the tail of the spatial power spectrum, a trend that can also be observed empirically in Fig. 2. Thus, a natural way to quantify the complexity of explanations is through a summary statistic that captures the tail behavior of the spatial power spectrum of the input gradient.

To this end, we introduce *Expected Frequency*, a simple yet effective summary statistic defined as:

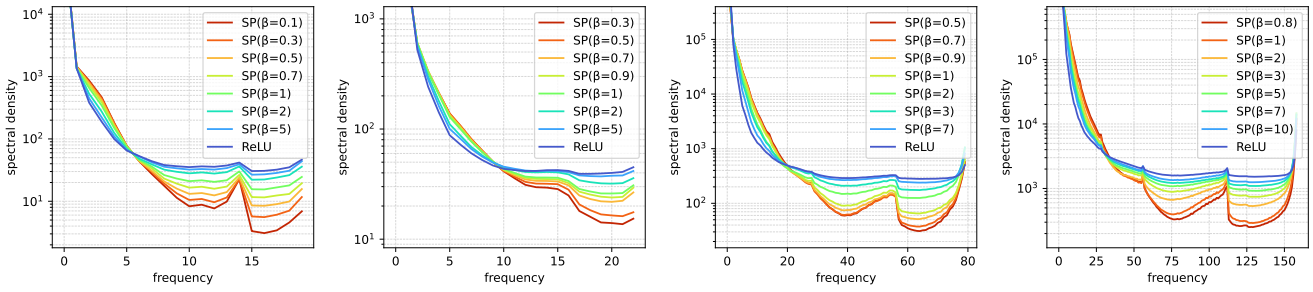
$$\mathrm{EF}(e_f(x)) \coloneqq \int \omega \, \mathrm{S}_{e_f(x)}(\omega) d\omega. \tag{1}$$

Here, S represents the spatial power spectrum of the explanation method $e_f : \mathbb{R}^n \to \mathbb{R}^n$ applied to the function f, which may simply be VanillaGrad. Hereafter, we may drop the subscript f from e_f for notational simplicity. As we focus on image data, we use the radial average in frequency domain to have a 1D power spectrum $\mathrm{S}(\omega)$.

Since the concept of high-frequency information is inherently relative, it can be measured in various ways. We found that expected frequency provides an effective statistic of model's reliance on high-frequency features. Nonetheless, the use of expectation is not pivotal for the results and is solely based on simplicity and practical effectiveness. The utility of EF is emphasized in the following remark:

Remark 1. *According to Eq. (1), Expected Frequency is influenced by both the model and the explanation method used. A lower EF value corresponds to smoother, or simpler, explanations in the spatial domain.*

We hypothetize that expected frequency is influenced by neural networks' dependence on high-frequency components for making predictions. It is crucial to note that EF captures the tail behavior of spatial power spectrum of



(a) (b) (c) (d)

Figure 2. **ReLU Affects Gradient TSPS via Network's TPS.** This figure presents a practical implication of our theory on the tail behavior of the spatial power spectrum in relation to function sharpness, as formalized in Lemma 1. The plots depict the TSPS of input-gradients of networks trained with *identical* architectures and hyperparameters, varying *only* in smoothness parameter of ReLU (indicated in the legend) and datasets: **(a)** Fashion MNIST, **(b)** CIFAR10, **(c)** Imagenette (112×122), and **(d)** Imagenette (224×224). Refer to Tab. 3 in the Appendix for details on the hyperparameters used. In all plots, reducing smoothness by increasing β (shown as SP(β) in the legends) results in power spectra with heavier tails, leading to more complex explanations. Since larger complexity necessitates removing more of the tail, this contributes to a larger explanation gap, defined in Eq. (7). Considering that not all frequencies are equally relevant for classification, notable spikes in mid-range frequencies are particularly evident in **(a)**, **(c)**, and **(d)**. Additionally, some spectra exhibit a jump at the extreme end of the tail (*i.e.*, very high frequencies). We hypothesize that this is due to the model's reliance on edges of the objects, which are less distinct in CIFAR10 but more pronounced in Fashion MNIST, explaining the absence of a sharp tail in **(c)** and its rise in other cases.

the gradient. In contrast, a network's utilization of high-frequency information is quantified by the tail of its power spectrum. In the next section, we establish a connection between the tail spatial power spectrum of the gradient and the tail of power spectrum of the network in data modalities with high input feature correlation.

3.3. TPS of a Network and TSPS of Gradient

Having established a measure for explanation complexity in the spatial domain, we now take steps toward addressing the question: *What causes the input gradient to look noisy?*

Our goal is to support an intuitive yet challenging-to-formalize claim: the presence of noise in the input gradient reflects the network's reliance on high-frequency information for its predictions. Ideally, if we had direct access to the network's Fourier transform, we could compute its power spectrum and directly compare it with the input gradient spectrum to verify this claim. However, since the Fourier transform of arbitrary networks is inaccessible, we instead have to look at the network's power spectrum through the aperture of the spatial power spectrum of its gradient.

This idea is captured in the following theorem, which establishes a relationship between the tails of two power spectra: the tail of the spatial power spectrum of the input gradient and the tail of the power spectrum of the network itself. For clarity and narrative coherence, we defer the proof and technical details to Appendix E.

Theorem 1 (informal)**.** *In data domains with high input feature correlation,* e.g. *image data; given a trained neural network $f(x)$, the tail behavior of the power spectrum of $f(x)$, is directly proportional to that of the spatial power spectrum of $\nabla f(x)$, denoted by $\mathrm{S}_{\nabla f}(\omega)$.*

Proof sketch. We note that the decay rate of the tail of the power spectrum of a function is influenced by the sharp changes in the function created by the network. We show that the decay rate depends on the existence of such sharp transitions rather than the number of their occurrences. Therefore, we consider a single training sample and a single explained sample with at least one sharp change. By applying a linear approximation around this sharp change in the spatial domain, we make the spatial Fourier transform analytically tractable; see Figs. 10 and 12 for a conceptual visualization, and Appendix E for technical details. □

In simple terms, assuming a high correlation between input features in the spatial domain (such as images), a heavier tail in the network's power spectrum leads to a heavier tail in the spatial power spectrum. Intuitively, a heavier tail indicates higher explanation complexity, leading to a higher expected frequency, as shown in Fig. 1 (see Appendix G).

A key aspect of Theorem 1 is the emphasis on the "tail". This terminology signifies that the high-frequency characteristics of the network—specifically, its tail behavior—are mirrored in the tail of the spatial power spectrum of the gradient. However, this correspondence may not accurately reflect the network's behavior at or near zero frequency.

In this section, we formalized an intuitive statement in Theorem 1. Next, we explore the practical implications of this theory, which also serve as an empirical validation of our insights. Specifically, we tweak the sharp transitions

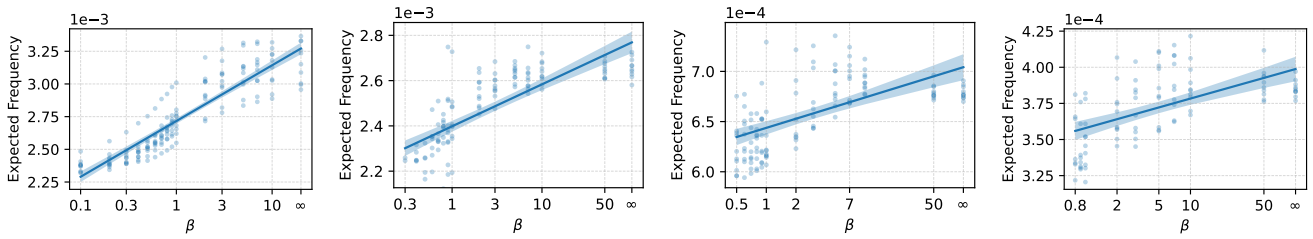


Figure 3. **Impact of Smoothness Parameter on Expected Frequency.** This figure shows EF, defined in Eq. (1), used as a summary statistic of the frequency content of explanations in the spatial domain with respect to the smoothness parameter β in a (log) x-axis. To obtain the 95% CI, we repeated experiments using 10 seeds for each dataset: **(a)** Fashion MNIST, **(b)** CIFAR-10, **(c)** Imagenette (122×122), and **(d)** Imagenette (224×224). The results consistently show that, on average, EF in the spatial domain increases with the smoothness parameter β, where $\beta \to \infty$ corresponds to a ReLU network. The variance observed in the scatter plots, we believe, is due to the weight initialization.

introduced by ReLU in the network and test whether our theoretical predictions align with empirical observations.

3.4. On the TPS of ReLU Networks

Having established the connection between the TPS of the network and the TSPS of its input gradient, we now turn to the first practical implication of our theory. Specifically, in this section, we control the TPS of a ReLU network and observe its impact on the TSPS of the input gradient.

Given its widespread use in modern architectures, and the growing evidence that ReLU induces sharp transitions in neural networks [19, 28, 47, 54], we focus on the effect of ReLU for studying its effect on the TPS of a network. Nonetheless, there are many other design decisions, see Appendix B, that can be studied in this regard.

To this end, we introduce the following lemma, which provides a simple approach to reduce the sharpness of transitions in the final trained function, thereby leading to a faster-decaying tail of the power spectrum of the network.

Lemma 1 (informal)**.** *Let ξ be a smoothed version of an activation function ϕ, achieved by convolving it with a Gaussian function with precision β,* i.e. *$\xi = \phi * g_\beta$. Let $f_\phi(x)$ and $f_\xi(x)$ denote the classifiers trained on $\mathcal{X}$ using activation functions ϕ and ξ, respectively. Then f_ϕ, exhibits a heavier tail in its power spectrum compared to f_ξ.*

Proof sketch. Building on prior work in kernel methods, we show that convolving ReLU with a Gaussian results in a smoother kernel without explicit characterization of the kernel. See Appendix F for formal details. □

To summarize, using the above lemma, we can control the TPS of the network, and through Theorem 1, we expect to observe its effect on the TSPS of the input gradient. Specifically, we employ a smooth parameterization of ReLU and gradually interpolate from a smoothed version toward the standard ReLU. This corresponds to increasing the precision parameter β of the Gaussian convolution, where ReLU is recovered in the limit $\beta \to \infty$. This process effectively interpolates between networks with smoother boundaries—exhibiting faster decay rates in their power spectrum—toward ReLU-like networks characterized by sharp transitions and heavier tails.

Based on our theory, we anticipate that as β increases, the TSPS of the input gradient will exhibit heavier tails, as illustrated in Fig. 2. This trend should also manifest in our measure of explanation complexity, namely, the expected frequency. Since a lower EF corresponds to simpler explanations, we expect smoother explanations as β decreases, which can also be seen in Fig. 1. The results in Fig. 3 confirm this, consistently showing that EF in the spatial domain increases with the smoothness parameter β of ReLU.

We would like to highlight that Lemma 1 is being leveraged in two ways: (1) to validate our theoretical prediction regarding the effect of the TPS of the network on the TSPS of the input gradient and (2) to demonstrate a practical application of our theory beyond merely quantifying explanation complexity—namely, controlling it. Additionally, Lemma 1 has a straightforward proof and is empirically verified inAppendix F. Please refer to Sec. 5 for implementation details and hyperparamters.

As a more general application of this framework, in the next section, we introduce measures for faithfulness and examine its trade-off to complexity as measured by EF.

4. The Complexity-Faithfulness Trade-off

Surrogate models used for explanations aim to reduce complexity, often at the cost of faithfulness. This trade-off raises concerns, as the disparity between explanations generated by the original and surrogate models can be arbitrarily large [57]. These methods rely on heuristically chosen hyperparameters, resulting in implicit surrogates with unknown properties, including their performance. This discrepancy can be significant enough to generate the same

explanations for different model behaviors.

Thus, it is crucial to formally define the explanation gap in order to bring the faithfulness-complexity trade-off to the attention of researchers in XAI. We introduce a unifying spectral framework quantifying both aspects systematically.

4.1. The Explanation Gap

Let $\tilde{f}$ denote a (typically implicit) surrogate model for a function f, generated by an explanation method e. We define the *explanation gap* as the squared L^2-norm of the difference between the gradients of f and $\tilde{f}$:

$$\mathcal{G}_{\mathcal{X}}(f, \tilde{f}) = \int_{x \in \mathcal{X}} \|\nabla f(x) - \nabla \tilde{f}(x)\|_2^2 dx \tag{2}$$

Here, the subscript $\mathcal{X}$ represents the training data used to learn f, emphasizing the dataset dependence of the measure. This dependence means that explanation gaps are *not* directly comparable across different datasets without appropriate normalization. For brevity, we may omit $\mathcal{X}$, though its influence remains implicit.

Notably, this measure exhibits two scaling behaviors: (1) it scales with the input dimension, and (2) it indirectly depends on the explanation method used to generate the surrogate $\tilde{f}$. As a result, different post-hoc explanation methods yield different explanation gaps.

Remark 2. *An immediate property of this definition is that when using VanillaGrad, which does not create a surrogate model (*i.e.*, $\tilde{f} = f$), we obtain $\mathcal{G} = 0$.*

Since explanation methods are typically designed and refined through trial and error rather than with the explicit goal of constructing smooth surrogate models, the resulting surrogates are often implicit and inaccessible. This makes direct measurement of the explanation gap challenging, necessitating the identification of a suitable proxy measure.

In Sec. 4.3, we express the explanation gap in the Fourier domain to establish a connection between the gap and the TPS of the network. This reformulation provides several benefits: (1) it offers deeper insights into the nature of the explanation gap, (2) it links the gap to our definition of expected frequency, and (3) it clarifies the impact of surrogates on the explanation gap.

To build the necessary intuition, we first briefly introduce explanation methods in the Fourier domain in Sec. 4.2. For a more detailed discussion, see Appendix A and [41].

4.2. Explanation Methods as Low-Pass Filters

Previous works have provided unifying definitions of explanation methods, including set-theoretic [16] and probabilistic [41] perspectives. We adopt the probabilistic view, as it is more suitable for a spectral analysis.

In this view, many gradient-based explanations can be written as:

$$e_f(x) = \mathbb{E}_{\tilde{x} \sim \mathcal{N}(x)}[\nabla f(\tilde{x})], \tag{3}$$

where $\tilde{x} \sim \mathcal{N}(x)$, a method-specific perturbation distribution. This form unifies a range of methods (see [41], App. C) and facilitates a spectral analysis of perturbations, particularly in the Fourier domain (see Appendix A).

To develop theoretical insights, we focus on VanillaGrad, corresponding to $\mathcal{N}(x) = \delta(x)$, *i.e.*, no perturbation. As a pure gradient method, it reveals the network's full spectral behavior without surrogate-induced filtering.

In the Fourier domain:

$$\nabla f(x) \propto \int \omega \widehat{f}(\omega), d\omega, \tag{4}$$

highlighting that, in ReLU networks, high-frequency components (in TPS), dominate the explanation (in TSPS), often resulting in noisy and grainy visualizations [41], particularly in ReLU networks.

By contrast, perturbation-based methods suppress high-frequency components, effectively acting as low-pass filters. This smoothing improves visual clarity but may compromise faithfulness by shifting explanations away from the original model.

In the next section, we build on this formalism to refine our definition of the explanation gap.

4.3. The Explanation Gap in Fourier Domain

Given that perturbations in explanation methods, or the creation of surrogates in general, are to suppress the high-frequency behavior of f, we can refine the definition in Eq. (2) to better account for this phenomenon. By applying Parseval's theorem—which states that the L^2-norm of a function is equal in both the time and Fourier domains—we can rewrite the gap as follows:

$$\begin{aligned} \mathcal{G}(f, \tilde{f}) &= \int_{x \in \mathcal{X}} \|\nabla f(x) - \nabla \tilde{f}(x)\|^2 dx \\ &\propto \int_{\omega \in \mathbb{R}} \omega^2 \|\widehat{f}(\omega) - \widehat{\tilde{f}}(\omega)\|^2 d\omega \end{aligned} \tag{5}$$

The above integral can be broken down into high and low frequencies for some threshold ω^*. Since surrogates suppress high-frequency components, one expects the integral on low frequency components to vanish, *i.e.* $\int_{\omega \in \mathcal{F}_{\text{low}}} \approx 0$. Therefore, we can approximate this as:

$$\mathcal{G}(f, \tilde{f}) \approx \int_{\omega \in \mathcal{F}_{\text{high}}} \omega^2 \|\widehat{f}(\omega) - \widehat{\tilde{f}}(\omega)\|^2 d\omega \tag{6}$$

This formulation emphasizes that the explanation gap is largely influenced by the high-frequency components of the classifier—specifically, *the model's reliance on high-frequency features*, as captured by the TPS of f. However, a familiar challenge arises once again: the network's power spectrum is not directly accessible, making it difficult to analyze its tail behavior.

Recall from Sec. 3 that we introduced expected frequency as a means to indirectly infer properties of the network's TPS through the TSPS of the input gradient. By leveraging this connection, we can now close the loop and use variations in EF as a proxy for the explanation gap, as detailed in Sec. 4.4.

4.4. Empirical Evaluation of the Explanation Gap

In the previous sections, we expressed the explanation gap in the Fourier domain, showing that it is primarily influenced by the TPS of the network. We also established that the TPS is reflected in the spatial power spectrum, $\mathrm{S}_{\nabla f}$, and introduced expected frequency as a measure of the network's reliance on high-frequency information, summarizing the TSPS of the input gradient. In this section, we bring these connections together to quantify the explanation gap.

More formally, we can use the absolute change in EF caused by an explanation method as a proxy for explanation gap, expressed as:

$$\mathcal{G}(f, \tilde{f}) \sim \Delta\,\mathrm{EF}(e_f) \coloneqq |\,\mathrm{EF}(\nabla f) - \mathrm{EF}(e_f)| \quad (7)$$

See Appendix E for derivations.

Remark 3. *Due to the inherent scaling properties of the gap, this quantity is primarily meaningful in a relative sense when comparing different explanation methods. Informally, it reflects the degree of engineering involved in creating a surrogate to achieve lower-complexity explanations.*

Thus, $\Delta\,\mathrm{EF}$ *reveals a range of behaviors rather than a dichotomy between post-hoc and ante-hoc explainability.*

This definition closely resembles a direct difference in the pixel domain, at the same time offers advantages such as shift invariance (via the Fourier transform) and rotation invariance (through the power spectrum). It also allows for flexible normalization strategies in the Fourier domain. While this measure can be highly correlated with direct pixel-based differences, its foundation in spectral analysis provides a more formal and unifying perspective on explanation complexity and faithfulness.

As demonstrated in [41], many post-hoc explanation methods function as low-pass filters, suppressing high-frequency model behavior through implicit surrogate models. Having introduced the explanation gap as a measure of faithfulness—supported theoretically in Theorem 1—we now use this proxy to quantify the gap introduced by various post-hoc explainability methods. However, this should not be interpreted as an endorsement of methods that suppress high-frequency components implicitly, as such alterations may distort the original model's behavior in ways that are unknown and difficult to quantify.

To quantify the explanation gap, we compute EF, defined in Eq. (1) and measure its change after introducing a surrogate, using this difference as a proxy measure, as in Eq. (7).

	Method	$\downarrow$ EF + $\downarrow \Delta$ EF	
		ReLU	SP($\beta = .9$)
Imagenette–CNN	VanillaGrad [53]	$.390 + \Delta.000$	$.202 + \Delta.000$
	SmoothGrad [54]	$.286 + \Delta.104$	$.196 + \Delta.005$
	IntGrad [58]	$.396 + \Delta.007$	$.205 + \Delta.003$
	GuidedBp [55]	$.300 + \Delta.090$	$.202 + \Delta.000$
	DeepLift [51]	$.394 + \Delta.005$	$.204 + \Delta.002$
	GradCAM [50]	$.293 + \Delta.097$	$.177 + \Delta.025$
	LRP [9]	$.394 + \Delta.005$	Undefined

Table 1. **The Continuous Range of Faithfulness-Complexity Trade-offs.** This table shows the evaluation of expected frequency in Eq. (1) and explanation gap in Eq. (7) on different post-hoc explanation methods, scaled by 10^4. As can be seen in this table, one can achieve smaller expected frequency while still having zero explanation gap (VanillaGrad row). Thus, Δ EF captures a continuous range of behaviors rather than enforcing a rigid distinction between post-hoc and ante-hoc explainability.

	Method	$\downarrow$ EF + $\downarrow \Delta$ EF	
		ResNet50[25]	ViT-B16[63]
ImageNet	VanillaGrad [53]	$.263 + \Delta.000$	$.222 + \Delta.000$
	SmoothGrad [54]	$.247 + \Delta.017$	$.221 + \Delta.001$
	IntGrad [58]	$.253 + \Delta.010$	$.221 + \Delta.001$
	GuidedBp [55]	$.294 + \Delta.031$	$.222 + \Delta.000$
	DeepLift [51]	$.254 + \Delta.009$	$.224 + \Delta.003$
	GradCAM [50]	$.133 + \Delta.130$	$.181 + \Delta.041$
	LRP [9]	$.282 + \Delta.019$	Undefined

Table 2. **The Faithfulness-Complexity Trade-off in Other Architectures.** This table presents the evaluation of the expected frequency from Eq. (1) and the explanation gap from Eq. (7) across various post-hoc explanation methods for two pretrained architectures on ImageNet, scaled by 10^4. As shown, GradCAM, a widely used method known for generating smoother explanations, consistently introduces the largest explanation gap across different architectures. In contrast, most classical saliency methods, originally designed for feedforward convolutional networks, exhibit lower variability when applied to vision transformers. Also the activation function used in VIT is GELU which exhibits a high empirical similarity with SP (but an empirical comparison between RELU and GELU ViT, shows that the ViT architecture contributes to lower variability more than the activation used see Tab. 4).

Note that the smoothness can be seen even in single images (Fig. 14), however, in Tabs. 1 and 2 we report our EF measure average over 1K images. We observe stds to be vanishingly small, and as such chose not to report it.

Thus, the explanation gap is computed by first determining EF for both VanillaGrad and a given explanation method, then taking the difference. Furthermore, to demonstrate the broader applicability of our approach, we measure

these quantities across different architectures, even in cases where our theoretical results may not strictly apply. The results are presented in Tabs. 1 and 2.

5. Implementation Details and Ablations

To accommodate both sharp and smooth ReLU variants within a Smooth Parameterization (SP), we used SoftPlus, an efficient approximation:

$$\text{SoftPlus}(x;\beta) = \frac{1}{\beta}\ln(1+e^{\beta x}) \approx \text{ReLU} * g_\beta(x) \quad (8)$$

where β controls smoothness, corresponding to the precision of the Gaussian convolution g_β.

To isolate ReLU's effects, we used feedforward convolutional networks, avoiding interference of confounding factors like residual connections and batch normalization (Appendix B). Experiments were conducted across four datasets with varying input sizes (Fig. 2).

Since ReLU networks typically converge faster than smooth parameterizations, we set dataset-specific validation accuracy caps for early stopping, preventing initialization biases (see Appendix B for an ablation and Fig. 5 for a comparison of validation accuracies achieved with each parameterization). This ensured meaningful comparisons under similar training budgets (see Tab. 3 for hparams).

For ImageNet [18], pretrained networks were used to measure frequency and explanation gaps (Tabs. 1 and 2), leveraging the Captum library [34].

We also conducted ablation studies on network depth and learning rate (Appendix B). Depth had little effect on spectral decay rates, while learning rate influenced curve shapes without altering overall tail behavior (Fig. 2).

Notably, differences in the spatial power spectra of various models are robust and often observable from a single image. However, for improved reliability, we averaged decay rates across 1K images, using a consistent batch for all activations and depths within each dataset. Furthermore, we have used normalized S in Eq. (1), to have a distribution over frequencies, yet the values are comparable without normalization within the same architecture and dataset.

Given that Fourier analysis depends on the distribution of gradient magnitudes, we employed the inverse transformation method to normalize rankings for each pixel [22, 56]. This ensured that results were independent of absolute gradient magnitudes, aligning with the intuition that explanation magnitudes can sometimes be uninformative and are better interpreted through rankings (see Appendix D).

6. Limitations and Future Work

We conjecture that expanding this research into a unified theoretical framework to analyze various design decisions would necessitate a more advanced mathematical theory of deep networks, an area that is currently lacking in the field [44]. Nonetheless, an comprehensive empirical analysis of the design choices can pave the way for a more advanced and encompassing theory.

Given that this work relies on kernel methods for the connection between the tail of the network's power spectrum and the tail of the spatial power spectrum of input-gradient, caution is advised, particularly where the kernel perspective becomes counterintuitive, as seen in the results on depth [7], or the assumption of infinite depth. Nonetheless, it can be seen empirically, particularly in our work, that the conclusions hold in finite width networks.

Our analysis assumes continuity in the spatial domain for analytical convenience. A suitable discretization of our framework is necessary for each resolution for insights into the tail behavior of networks in practice.

We believe this work lays the foundation for a more formal analysis and a systematic investigation of potential approaches, such as neural architecture search using the spectrum's tail as an optimization objective.

Additionally, it presents a novel perspective on the traditional ante-hoc vs. post-hoc explainability dilemma, reframing them as two orthogonal axes (EF $+\Delta$ EF) with a continuum of possible behaviors.

7. Conclusion

In this work, we introduced a unifying spectral framework to systematically analyze the complexity-faithfulness trade-off in gradient-based explainability. By leveraging Expected Frequency (EF) as a principled metric, we quantified a network's reliance on high-frequency information and established a connection between network behavior and explanation complexity. Our findings demonstrate that explanation complexity and faithfulness are deeply intertwined through spectral properties, with surrogate models often introducing a significant "explanation gap."

Through theoretical analysis and empirical validation, we demonstrated that controlling the spectral properties of ReLU networks can lead to smoother explanations without compromising faithfulness. Our results offer practical implications for designing both neural networks and explainability methods that maintain interpretability with higher faithfulness to the original model. This study lays the foundation, summarized in Fig. 4, for future research in more formal approaches to explainability of neural networks with respect to their inherent explainability properties.

Acknowledgements

This project is partially supported by Region Stockholm through MedTechLabs, and Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Scientific computation was enabled by the

supercomputing resource Berzelius, provided by the National Supercomputer Center at Linköping University and the Knut and Alice Wallenberg foundation.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps, 2020. arXiv:1810.03292 [cs, stat]. 1, 12

[2] Genevera I. Allen. Automatic Feature Selection via Weighted Kernels and Regularization. *Journal of Computational and Graphical Statistics*, 22(2):284–299, 2013. 14

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, 2015. 2

[4] David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The Shattered Gradients Problem: If resnets are the answer, then what is the question?, 2018. arXiv:1702.08591 [cs, stat]. 12

[5] Daniel Beaglehole, Ioannis Mitliagkas, and Atish Agarwala. Feature learning as alignment: a structural property of gradient descent in non-linear neural networks, 2024. arXiv:2402.05271 [cs, stat]. 14, 16

[6] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and Aggregating Feature-based Model Explanations, 2020. arXiv:2005.00631 [cs]. 1, 2, 3

[7] Alberto Bietti and Francis Bach. Deep Equals Shallow for ReLU Networks in Kernel Regimes, 2021. arXiv:2009.14397 [stat]. 8, 12, 15

[8] Alberto Bietti and Julien Mairal. On the Inductive Bias of Neural Tangent Kernels, 2019. arXiv:1905.12173. 14, 16

[9] Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers, 2016. arXiv:1604.00825 [cs]. 7, 14

[10] Kirill Bykov, Anna Hedström, Shinichi Nakajima, and Marina M.-C. Höhne. NoiseGrad: Enhancing Explanations by Introducing Stochasticity to Model Weights, 2022. arXiv:2106.10185 [cs]. 2, 3

[11] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos Networks: Alignment is All We Need for Interpretability, 2022. 2

[12] Zhicheng Cai, Hao Zhu, Qiu Shen, Xinran Wang, and Xun Cao. Towards the Spectral bias Alleviation by Normalizations in Coordinate Networks, 2024. arXiv:2407.17834 [cs]. 12

[13] Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. Can I Trust the Explainer? Verifying Post-hoc Explanatory Methods, 2019. arXiv:1910.02065 [cs]. 2

[14] Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. True to the Model or True to the Data?, 2020. arXiv:2006.16234 [cs, stat]. 1

[15] Jianbo Chen, Mitchell Stern, Martin J. Wainwright, and Michael I. Jordan. Kernel Feature Selection via Conditional Covariance Minimization, 2018. arXiv:1707.01164 [stat]. 14

[16] Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by Removing: A Unified Framework for Model Explanation, 2022. arXiv:2011.14878 [cs, stat]. 6

[17] Amit Daniely, Roy Frostig, and Yoram Singer. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity, 2017. arXiv:1602.05897. 14

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. ISSN: 1063-6919. 8

[19] Matteo Gamba, Hossein Azizpour, and Marten Bjorkman. On the lipschitz constant of deep networks and double descent. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. 1, 5

[20] Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Ronen Basri. On the Similarity between the Laplace and Neural Tangent Kernels, 2020. 14, 16, 20

[21] Magda Gregorová, Jason Ramapuram, Alexandros Kalousis, and Stéphane Marchand-Maillet. Large-scale Nonlinear Variable Selection via Kernel Random Features, 2018. arXiv:1804.07169 [cs]. 14

[22] Modris Greitans. Spectral analysis based on signal dependent transformation. *Proc. SMMSP 2005*, pages 179–184, 2005. 8, 17

[23] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post Hoc Explanations, 2022. arXiv:2206.01254 [cs]. 1

[24] Kosuke Haruki, Taiji Suzuki, Yohei Hamakawa, Takeshi Toda, Ryuji Sakai, Masahiro Ozawa, and Mitsuhiro Kimura. Gradient Noise Convolution (GNC): Smoothing Loss Function for Distributed Large-Batch SGD, 2019. arXiv:1906.10822 [cs, stat]. 12

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, 2015. arXiv:1512.03385 [cs]. 7

[26] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A Benchmark for Interpretability Methods in Deep Neural Networks, 2019. arXiv:1806.10758 [cs, stat]. 2

[27] Zejiang Hou and S. Y. Kung. A Kernel Discriminant Information Approach to Non-linear Feature Selection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2019. ISSN: 2161-4407. 14

[28] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks, 2020. arXiv:1806.07572. 1, 5, 13

[29] Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations, 2024. arXiv:2310.02557 [cs]. 2, 23

[30] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences, 2018. arXiv:1807.02582 [cs, stat]. 14

[31] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided Integrated Gradients: An Adaptive Path Method for Removing Noise, 2021. arXiv:2106.09788 [cs]. 2

[32] Beomsu Kim, Junghoon Seo, SeungHyun Jeon, Jamyoung Koo, Jeongyeol Choe, and Taegyun Jeon. Why are Saliency Maps Noisy? Cause of and Solution to Noisy Saliency Maps, 2019. arXiv:1902.04893 [cs, stat]. 2

[33] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (Un)reliability of saliency methods, 2017. arXiv:1711.00867 [cs, stat]. 1

[34] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch, 2020. 8

[35] Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. Cartoon Explanations of Image Classifiers, 2022. arXiv:2110.03485 [cs]. 2

[36] Stefan Kolek, Robert Windesheim, Hector Andrade Loarca, Gitta Kutyniok, and Ron Levie. Explaining Image Classifiers with Multiscale Directional Image Representation, 2023. arXiv:2211.12857 [cs]. 2

[37] Gabriel Laberge, Yann Pequignot, Foutse Khomh, Mario Marchand, and Alexandre Mathieu. Partial order: Finding Consensus among Uncertain Feature Attributions, 2023. arXiv:2110.13369 [cs]. 17

[38] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. Human Evaluation of Models Built for Interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7:59–67, 2019. 2

[39] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 17

[40] Giovanni Luca Marchetti, Christopher Hillar, Danica Kragic, and Sophia Sanborn. Harmonics of Learning: Universal Fourier Features Emerge in Invariant Networks, 2024. arXiv:2312.08550 [cs]. 2, 23

[41] Amir Mehrpanah, Erik Englesson, and Hossein Azizpour. On Spectral Properties of Gradient-Based Explanation Methods. In *Computer Vision – ECCV 2024*, pages 282–299, Cham, 2025. Springer Nature Switzerland. 1, 6, 7, 12, 13

[42] Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montufar. Characterizing the Spectrum of the NTK via a Power Series Expansion, 2023. arXiv:2211.07844 [cs]. 21

[43] Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and Generalization in Neural Networks: an Empirical Study, 2018. arXiv:1802.08760 [cs, stat]. 19

[44] Philipp Petersen and Jakob Zech. Mathematical theory of deep learning, 2024. arXiv:2407.18384 [cs]. 8

[45] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the Spectral Bias of Neural Networks, 2019. arXiv:1806.08734 [cs, stat]. 2, 3

[46] José Ribeiro, Lucas Cardoso, Vitor Santos, Eduardo Carvalho, Níkolas Carneiro, and Ronnie Alves. How Reliable and Stable are Explanations of XAI Methods?, 2024. arXiv:2407.03108 [cs]. 1

[47] Mihaela Rosca, Theophane Weber, Arthur Gretton, and Shakir Mohamed. A case for new neural network smoothness constraints, 2021. arXiv:2012.07969 [cs, stat]. 1, 5

[48] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 2

[49] Jessica Schrouff, Sebastien Baur, Shaobo Hou, Diana Mincu, Eric Loreaux, Ralph Blanes, James Wexler, Alan Karthikesalingam, and Been Kim. Best of both worlds: local and global explanations with human-understandable concepts, 2022. arXiv:2106.08641 [cs]. 2

[50] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. arXiv:1610.02391 [cs]. 2, 7, 14

[51] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences, 2017. arXiv:1605.01713 [cs]. 2, 7, 14

[52] James B. Simon, Sajant Anand, and Michael R. DeWeese. Reverse Engineering the Neural Tangent Kernel, 2022. arXiv:2106.03186 [cs]. 14, 20

[53] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, 2014. arXiv:1312.6034 [cs]. 2, 7, 14

[54] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise, 2017. arXiv:1706.03825 [cs, stat]. 1, 2, 3, 5, 7, 14

[55] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net, 2015. arXiv:1412.6806 [cs]. 2, 7, 14

[56] Petre Stoica and Randolph L. Moses. *Spectral analysis of signals*. Pearson Prentice Hall Upper Saddle River, NJ, 2005. 8, 17

[57] Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. arXiv:1710.08864 [cs]. 1, 5

[58] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, 2017. arXiv:1703.01365 [cs]. 2, 7, 14

[59] Yusuke Tsuzuku and Issei Sato. On the Structural Sensitivity of Deep Convolutional Networks to the Directions of Fourier Basis Functions, 2019. arXiv:1809.04098 [cs]. 2, 3

[60] Kristoffer Wickstrøm, Marina Marie-Claire Höhne, and Anna Hedström. From Flexibility to Manipulation: The Slippery Slope of XAI Evaluation, 2024. arXiv:2412.05592 [cs]. 1

[61] Maksymilian Wojtas and Ke Chen. Feature Importance Ranking for Deep Learning, 2020. arXiv:2010.08973 [cs]. 17

[62] Matthew A. Wright and Joseph E. Gonzalez. Transformers are Deep Infinite-Dimensional Non-Mercer Binary Kernel Machines, 2021. arXiv:2106.01506 [cs]. 14

[63] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual Transformers: Token-based Image Representation and Processing for Computer Vision, 2020. arXiv:2006.03677 [cs]. 7, 14

[64] Mengjiao Yang and Been Kim. Benchmarking Attribution Methods with Relative Feature Importance, 2019. arXiv:1907.09701 [cs, stat]. 2