On Spectral Properties of Gradient-based Explanation Methods

Amir Mehrpanah[1], Erik Englesson[1], and Hossein Azizpour[1]

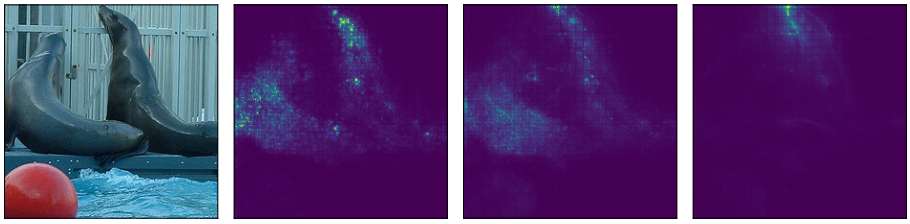
Department of Computer Science, Royal Institute of Technology, Stockholm, Sweden
{amirme,engless,azizpour}@kth.se

Abstract. Understanding the behavior of deep networks is crucial to increase our confidence in their results. Despite an extensive body of work for explaining their predictions, researchers have faced reliability issues, which can be attributed to insufficient formalism. In our research, we adopt novel probabilistic and spectral perspectives to formally analyze explanation methods. Our study reveals a pervasive spectral bias stemming from the use of gradient, and sheds light on some common design choices that have been discovered experimentally, in particular, the use of squared gradient and input perturbation. We further characterize how the choice of perturbation hyperparameters in explanation methods, such as SmoothGrad, can lead to inconsistent explanations and introduce two remedies based on our proposed formalism: (i) a mechanism to determine a standard perturbation scale, and (ii) an aggregation method which we call SpectralLens. Finally, we substantiate our theoretical results through quantitative evaluations.

Keywords: Probabilistic Machine Learning · Gradient-based Explanation Methods · Probabilistic Pixel Attribution Techniques · Explainability · Deep Neural Networks · Spectral Analysis

1 Introduction

Ensuring the interpretability of machine learning models is crucial for user's trust and understanding [2, 6, 45]. While considerable efforts have been dedicated to advancing explainability in the field [37, 47, 49, 51, 57, 62], there is a notable gap in systematically formalizing and consolidating existing knowledge [5, 9, 16, 19]. Insufficient formal analysis leaves too much room for different interpretations of the explanations, making them unreliable [27] which is detrimental to the very use case of explanations, i.e., human-critical applications. Therefore, more attention is essential to retrospectively analyze and structure the wealth of work done, creating a cohesive foundation for explainability [39, 61].
One main reason for the unreliability of explanation methods is their susceptibility to hyperparameters, leading to inconsistent explanations [4, 13, 33, 36, 48], see Fig. 1. Among the several works that formally study explanation methods [19, 37, 57], to our knowledge, none shed light on such inconsistencies. Therefore, in this work, we adopt a novel probabilistic and spectral perspective to characterize the observed inconsistencies and propose preliminary solutions.



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

Fig. 1: The figure illustrates sample explanations generated by varying hyperparameters of SmoothGrad-Squared [30, 55] for ResNet-50 on ImageNet (see Fig. 16 in the Appendix for more examples). It is evident from the figure that using different hyperparameters results in diverse explanations. Such variations have been demonstrated in recent studies for other explanation techniques as well [4, 13, 33, 48].

Our Contribution. We iterate our contributions in three parts:

- **Analysis:** We introduce a probabilistic representation for post-hoc gradient-based explanation methods in Sec. 2.1, and perform a spectral analysis of them in Sec. 3.1.
- **Identified Issues:** Our examination reveals a shared spectral bias stemming from the fundamental assumption of utilizing gradients to measure feature contributions. We show that the combined spectral effects of gradient and perturbation lead to band-pass filters modulating the explanations, which could lead to inconsistencies in the explanations, discussed in Sec. 3.2. Moreover, we conduct a spectral analysis of a few specific perturbation distributions that have been proposed in the literature in Sec. 3.3.
- **Solutions and Evaluations:** To address inconsistencies, we investigate two straightforward solutions that are finding an optimal perturbation scale based on the cosine similarity of the perturbation and the classifier in the frequency domain in Sec. 4.1, and an aggregation mechanism named SpectralLens in Sec. 4.2. Finally, we evaluate the proposed methods using the pixel removal strategy [48], acknowledging their effectiveness and limitations under Sec. 5.

2 Background

Notation. We consider a classification setting, that is, modeling $p(y|\boldsymbol{x})$ for labels $y \in \{1, \dots, k\}$ and $\boldsymbol{x} \in \mathbb{R}^n$ the set of input images. We denote the $\log p(y|\boldsymbol{x})$ by f. We denote standard Fourier bases by $\{\psi(\boldsymbol{\omega}, \boldsymbol{x})\}_{\boldsymbol{\omega} \in \mathbb{R}^n}$, according to which, under some regularity conditions [43], f can be decomposed as $f(\boldsymbol{x}) = \int \widehat{f}(\boldsymbol{\omega})\psi(\boldsymbol{\omega}, \boldsymbol{x})d\boldsymbol{\omega}$. In this representation, $\widehat{f}$ denotes the Fourier transform (FT) of f. We denote the power spectral density (PSD) of f by $\mathrm{S}_f(\boldsymbol{\omega}) = |\widehat{f}(\boldsymbol{\omega})|^2$.

2.1 A Probabilistic Representation of Explanations

In this view, each explanation method consists of three components: (1) a distribution for perturbation of the input, (2) an explainer, and (3) a summary statistic. We use $p(\tilde{\boldsymbol{x}}|\boldsymbol{x}, \sigma)$ to denote the distribution of perturbations of the input sample parameterized by σ, referred to as the perturbation scale. A definition of "explanation" is recovering the contribution of features to a prediction, which emanates from human's natural counterfactual reasoning [19]. Therefore, the explainer is usually the gradient with respect to input $\nabla_{\boldsymbol{x}}$, or approximations of it, *e.g.* the finite difference, given by $\Delta_{\boldsymbol{x}} f(\boldsymbol{x}) = \frac{f(\tilde{\boldsymbol{x}}) - f(\boldsymbol{x})}{\tilde{\boldsymbol{x}} - \boldsymbol{x}}$ (see Appendix D). Finally, we assume that the summary statistic is the expectation of the outputs, or the squared outputs, of the explainer, *e.g.* $(\nabla f(\tilde{\boldsymbol{x}}))^2$, over the chosen perturbation distribution, $p(\tilde{\boldsymbol{x}}|\boldsymbol{x}, \sigma)$, denoted as $\mathbb{E}_{p(\tilde{\boldsymbol{x}})}\left[(\nabla f(\tilde{\boldsymbol{x}}))^2\right]$, where the operations are carried out pixelwise. This is referred to as the explanation of f at $\boldsymbol{x}$ with a hyperparameter σ. This probabilistic view can, in fact, unify different explanation methods into a single form (see Appendix C). We provide theoretical justifications for why using squared gradients is preferable to gradients, explaining the empirical findings [30], and will discuss both cases in Sec. 3.1.

3 Spectral Analysis of Explanations

This section demonstrates the spectral property in explanations when using the gradient explainer and its interaction with perturbations, where a significant difference in the frequency of learned features leads to a selective attribution of features. In the following sections, we draw on insights from spectral analysis of signals and systems theory, utilizing this viewpoint to gain a deeper understanding of explanation methods. We demonstrate a spectral bias related to gradient-based explanation approaches, identify its source, and interpret various methods in the literature as efforts to alleviate it through the incorporation of low-pass filters, either explicitly [55, 57] or implicitly [42, 62]. We illustrate that the inconsistencies in explanation techniques stem from different strategies for mitigating the spectral bias. Finally, we suggest two straightforward approaches to tackle these inconsistencies: determining an optimal perturbation scale in Sec. 4.1 and aggregating the explanations in Sec. 4.2.

Writing a spectral representation for prediction, as shown in Appendix A.1, the model output with a perturbation distribution can be written as a convolution between the function and the perturbation distribution $f * p$, therefore we may use "perturbation kernel" and "perturbation distribution" interchangeably.

3.1 Spectral Representation of Explainers

Here, by drawing a connection between explainability techniques and the literature on spectral signal processing [12, 38, 56], we take a novel avenue for our analysis. This connection enables us to have a spectral representation of the explanations, as discussed in the next sections (see Appendix A for the derivations). It is important to note that the Fourier bases in our work is 1D and

per pixel, and should not be confused with the traditional Fourier bases used in image processing literature that expands along spatial dimensions.

Gradients in Spectral Domain. For explanation methods that use gradients and a perturbation kernel, $p(\tilde{\boldsymbol{x}})$ we have the following (see Appendix A.2).

$$\mathbb{E}_{p(\tilde{\boldsymbol{x}})}[\nabla f(\tilde{\boldsymbol{x}})] = \int i2\pi\boldsymbol{\omega}\widehat{f}(\boldsymbol{\omega})\widehat{p}(\boldsymbol{\omega})d\boldsymbol{\omega} \tag{1}$$

$$= -2\pi\left[\int \boldsymbol{\omega}(\widehat{f}_E(\boldsymbol{\omega})\widehat{p}_O(\boldsymbol{\omega}) + \widehat{f}_O(\boldsymbol{\omega})\widehat{p}_E(\boldsymbol{\omega}))d\boldsymbol{\omega}\right] \tag{2}$$

where $\widehat{p}$ denotes the FT of the perturbation kernel. We used $\widehat{f}_E$ and $\widehat{f}_O$ to denote a function $\widehat{f}$ is decomposed into real even and imaginary odd components $\widehat{f} = i\widehat{f}_O + \widehat{f}_E$[1]. The negative sign suggests that in case of using the gradients, we need to negate the actual values to obtain rankings. We will elaborate more on the negation of the attributions in Sec. 5.

Remark 1. Note the symmetry in the form of the spectral representation of the gradients. This form implies that it is inevitable for any perturbation kernel to extract asymmetric information about even and odd parts of the classifier f. Therefore, the gradient explainer with a kernel symmetric about the sample being explained, only extracts information about the odd parts of the classifier. This is problematic because the attribution is expected to reflect the total contribution to the prediction.

We will use this interesting symmetry later in Sec. 3.3 for analysis of a specific perturbation kernel.

Squared Gradients in Spectral Domain. On the other hand, for explanation methods which use squared gradients, we have the following (see Appendix A.3).

$$\mathbb{E}_{p(\tilde{\boldsymbol{x}})}\left[(\nabla f(\tilde{\boldsymbol{x}}))^2\right] = 4\pi^2 \int \|\boldsymbol{\omega}\|^2 \, \mathrm{S}_f(\boldsymbol{\omega}) \, \mathrm{S}_{\sqrt{p}}(\boldsymbol{\omega}) d\boldsymbol{\omega} \tag{3}$$

where S_f and $\mathrm{S}_{\sqrt{p}}$ denote the PSD of the model and square root of the perturbation kernel respectively.

We discuss the benefits of using squared gradients as in Eq. (3) over gradients as in Eq. (2) later in Sec. 3.3. However, it can be noticed here that the squared gradients, unlike gradients, are always non-negative, extract information about the PSD of f, and do not have a complex symmetric structure.

[1] Also note that as we are using real valued functions, other components *i.e.* real odd and imaginary even, are zero.

3.2 Gradient Explainers with Perturbations Are Band-Pass Filters

In our probabilistic framework, various explanation methods can be decomposed into three components (1) an explainer, (2) a perturbation kernel and (3) a summary statistic. In this section, we analyze the spectral effects of the two core components, namely gradient and perturbation. Importantly, we show that gradient acts as a high-pass filter, and perturbation acts as a low-pass filter. Before that, we need a more formal definition for a high-pass and low-pass filter that suits our study. Assuming that $\widehat{m}(\boldsymbol{\omega})$ denotes the FT of a perturbation kernel, we define low- and high-pass as follows.

Definition 1. *A filter $\widehat{m}(\boldsymbol{\omega})$ is low pass, if for a given $M > 0$ there exists a threshold $\boldsymbol{\omega}^*$ such that for any $\|\boldsymbol{\omega}\| > \|\boldsymbol{\omega}^*\|$, we have $\|\widehat{m}(\boldsymbol{\omega})\| < M$, and it is high pass if such a condition holds for $\|\boldsymbol{\omega}\| < \|\boldsymbol{\omega}^*\|$.*

For better readability, the proofs for the theorems and propositions are outlined in Appendix B, and we provide intuition as we progress.

Theorem 1. *Any perturbation kernel $p(\tilde{\boldsymbol{x}})$ mitigates the attribution of high-frequency features; simply put,* ***any perturbation is a low-pass filter****.*

The intuition is that any modification to a signal, would only remove information, and, as the intensity of perturbation increases, the removal goes from high-frequency components toward lower frequencies.

Theorem 2. *The gradient operator amplifies the attribution of high-frequency features; simply put,* ***the gradient operator is a high-pass filter****.*

From the stated theorems, one can see that a gradient-based explanation with perturbation creates a band-pass filter, whose shape is controlled by the perturbation hyperparameters (see Figs. 2 and 3).

Remark 2. This amplification-mitigation trade-off can be seen in several explanation methods, where perturbation kernels (low-pass filters) are utilized to mitigate the amplification effect (the high-pass filter) caused by the gradient. Therefore, the combined effect of the two filters usually forms **a band-pass filter** (see Appendix C for a non-comprehensive list).

3.3 Spectral Analysis of Specific Perturbation Kernels

In this section, we analyze the properties of special kernels used in the literature with remarks or propositions that justify, and thus shed light on, previous work's experimental observations. In most cases, we expand upon the squared gradients, Eq. (3), due to its analytical simplicity.

VanillaGrad via a Dirac Kernel. The very naive explanation method can be obtained with a Dirac kernel $p(\tilde{\boldsymbol{x}}) = \delta_{\boldsymbol{x}}(\tilde{\boldsymbol{x}})$ known as **VanillaGrad** (VG). Analyzing the squared version of VG and following Eq. (3), we know that the FT of the Dirac kernel is $\mathrm{S}_{\sqrt{p}}(\boldsymbol{\omega}) = \mathbf{1}(\boldsymbol{\omega})$, where $\mathbf{1}(\boldsymbol{\omega})$ is the constant function mapping all values to 1. Thus, it can be seen that we recover the scaled PSD of f, *i.e.* $\|\boldsymbol{\omega}\|^2 \mathrm{S}_f(\boldsymbol{\omega})$.

Remark 3. Scaling $\widehat{f}$ by frequency can be interpreted as the bias of the gradient explainer towards exaggeration of the attribution of high-frequency features [31]. Beside other justifications for gradient sparsity [8], spectral representation of VG^2 and domination of high-frequency features give an orthogonal *justification for why VanillaGrad explanations being usually scattered and noisy* (see Fig. 12).

SmoothGrad via a Gaussian Kernel. Letting $p(\tilde{\boldsymbol{x}}) = \mathcal{N}(\boldsymbol{x}, \sigma^2)$, *i.e.* a Gaussian with variance σ^2, centered on the sample being explained, we recover variants of an explanation method known as SmoothGrad (SG) or the squared gradient version SmoothGrad-Squared (SG^2) [30, 55], we can rewrite Eq. (3) as follows:

$$\mathbb{E}_{p(\tilde{\boldsymbol{x}})}\left[(\nabla f(\tilde{\boldsymbol{x}}))^2\right] \propto \int \|\boldsymbol{\omega}\|^2 \mathrm{S}_f(\boldsymbol{\omega}) e^{-8\pi^2\sigma^2\|\boldsymbol{\omega}\|^2} d\boldsymbol{\omega} \tag{4}$$

This shows that the contributions $\mathrm{S}_f(\boldsymbol{\omega})$ are modulated by the combined spectral effect of the Gaussian perturbation kernel and gradient operator, which is a band pass filter of the form $\|\boldsymbol{\omega}\|^2 e^{-8\pi^2\sigma^2\|\boldsymbol{\omega}\|^2}$, whose shape is controlled by the kernel variance (see Fig. 2). The explanations could be inconsistent, but generally become smoother as σ increases, which is a result of the shrinkage of the attribution of high-frequency features that can dominate the explanations with a small σ (see Fig. 1).

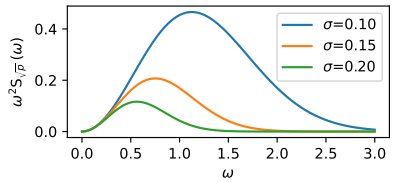


Fig. 2: A visualization of frequency bands, where $\mathrm{S}_{\sqrt{p}}(\omega)$ is defined according to Eq. (4), show that change in the noise scale σ leads to band-pass filters with different modes, hence, representing information of different frequencies.

When generating explanations, one is tempted to look for hyperparameters that make the explanation reflect what they *think it should* [16]. However, monitoring explanations from only one frequency band is potentially misleading and is likely to contradict itself simply by changing hyperparameters (see Fig. 5) [27,33].

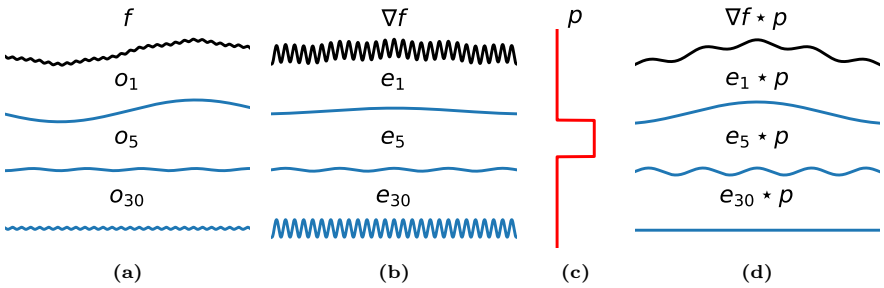


Fig. 3: A 1D visualization of amplification (b) and mitigation (d) effects caused by gradient and perturbation, respectively. For simplicity, we have shown a rectangular pulse (c) convolved with ∇f, but the kernel shape is usually more complex in practice, an example shown in Fig. 2. Moreover, functions in blue are the Fourier representation of the black functions. Due to the gradient operator, small high-frequency jitters in the decision boundary (a) amplify proportionally to their frequency (b). To remove the effect of high-frequency components of the decision boundary that usually dominate the explanations, researchers have proposed different perturbation kernels, visualized in (d). Theorem 2 corresponds to (a $\rightarrow$ b) and Theorem 1 corresponds to (b $\rightarrow$ d).

We formalize this statement, also known as the Rashomon effect (see Appendix B for the proof).

Proposition 1. *The SmoothGrad-Squared pixel attribution can contradict itself by changing the hyperparameters. Simply,* ***the Rashomon effect can occur in the SmoothGrad-Squared explanations****.*

Intuition. This proposition suggests a reason, *i.e.* removing different frequency bands, for inconsistencies we encounter in SG[2] explanations, which has been unknown so far [48]. Note that this occurs with different kernels, but we have chosen the one that leads to a simpler proof.

IntegratedGrad via a Rect Kernel. Setting the perturbation kernel to $p(\tilde{\boldsymbol{x}}) = \frac{1}{\sigma|\boldsymbol{s}-\boldsymbol{x}|}\,\mathrm{Rect}(\frac{\tilde{\boldsymbol{x}}-\boldsymbol{x}}{\sigma(\boldsymbol{s}-\boldsymbol{x})} - \frac{1}{2})$[2], where $\boldsymbol{s}$ is a chosen baseline image, we recover the perturbation kernel of a well known method called **IntegratedGrad** (IG) [57]. We use the spectral representation of gradients in Eq. (2), for a proposition about the gradient signs.

$$\mathbb{E}_{p(\tilde{\boldsymbol{x}})}[\nabla f(\tilde{\boldsymbol{x}})] \propto \int i\widehat{f}(\boldsymbol{\omega})\sin(\pi\sigma(\boldsymbol{s}-\boldsymbol{x})\boldsymbol{\omega})e^{i\pi\boldsymbol{\omega}\boldsymbol{r}}d\boldsymbol{\omega} \tag{5}$$

Where $\boldsymbol{r} = (1+2\frac{\boldsymbol{x}}{\sigma(\boldsymbol{s}-\boldsymbol{x})})$, appears because the kernel is not centered on the sample being explained. Note that the Fourier operator linearly expands along each dimension, but we use vector forms for notational simplicity. Furthermore, to

[2] Rect function is defined as $\mathrm{Rect}(\tilde{\boldsymbol{x}}) = \frac{1}{2}\,\mathrm{sign}(\frac{1}{2} - |\tilde{\boldsymbol{x}}|) + \frac{1}{2}$

study the effect of elementwise multiplication on the explanation quantitatively, we use **IG** to denote integrated gradients without the elementwise multiplication with input and **XIG** to denote the version with elementwise multiplication.

In explanation visualizations, the negative gradient sign has been used to show a negative contribution to the prediction [7, 53], to our knowledge it has never been formally justified. In the next proposition, we state that the negative gradient sign can be the direct result of the chosen perturbation kernel, formally shown in Appendix B.

Proposition 2. *The gradient sign in IG is affected by the perturbation level.*

This proposition entails that the negative sign in the gradient explainer does not necessarily reflect a negative contribution to the prediction, nor the positive sign necessarily reflects a positive contribution (see Figs. 13 and 14).

Remark 4. Since we assume continuity wrt σ, in our analysis, by Bolzano's theorem [50], there are hyperparameters that set the attribution to zero. Therefore, the gradient explainer may not produce a useful ranking compared to that of the gradient squared.

Remark 5. Proposition 2 also shows that adding explanations with different hyperparameters may cause destructive interference when the different values of $\boldsymbol{\omega}\widehat{p}(\boldsymbol{\omega})$ are out of phase. Hence, *squared gradients is a more appropriate design choice*, unless the use of phase information is planned as part of the method. This has been discovered experimentally [30] and confirmed in Sec. 5.

4 Proposed Solutions Based on the Spectral Analysis

4.1 Finding an Optimal Perturbation Scale

Note that the squared gradients in spectral domain (Eq. (3)) can be formulated as an inner product between the scaled PSD of the perturbation distribution and the PSD of the classifier *i.e.* $\mathbb{E}_{p(\tilde{\boldsymbol{x}})}\left[(\nabla f(\tilde{\boldsymbol{x}}))^2\right] = 4\pi^2\langle \mathrm{S}_f, \|\boldsymbol{\omega}\|^2\,\mathrm{S}_{\sqrt{p}}\rangle$. Assuming unimodality, a nonzero contribution, and that the norm $\left\|\|\boldsymbol{\omega}\|^2\,\mathrm{S}_{\sqrt{p}}\right\|$ exists, then the cosine similarity is maximized at a certain perturbation scale that has the highest similarity with S_f. A higher cosine similarity guarantees recovering more information about the classifier.

Definition 2. *The cosine similarity between the PSD of the classifier and that of the perturbation kernel given by*

$$\textit{Similarity}(\mathrm{S}_f, \|\boldsymbol{\omega}\|^2\,\mathrm{S}_{\sqrt{p}}) = 4\pi^2\frac{\langle \mathrm{S}_f, \|\boldsymbol{\omega}\|^2\,\mathrm{S}_{\sqrt{p}}\rangle}{\|\mathrm{S}_f\|\,\left\|\|\boldsymbol{\omega}\|^2\,\mathrm{S}_{\sqrt{p}}\right\|} \tag{6}$$

This quantity can be used as a way to determine the optimal perturbation scale per pixel to maximize the information extracted with the squared gradients. Note that, the word "optimal" refers to extracting the most information by the inner product, *whether deemed relevant by a human or not.*.

For one of the perturbation kernels we have presented earlier in Sec. 3.3, we have optimized this quantity on multiple levels, such as dataset level Fig. 4, image level Tab. 1, and, pixel level Fig. 5d. Finally, we have experimentally verified the effectiveness of this approach in Sec. 5.

Computing the Similarity of Classifier and Gaussian Kernel. To compute the similarity, we first need to evaluate the kernel norm as follows

$$\| \|\boldsymbol{\omega}\|^2 e^{-8\pi^2\sigma^2\|\boldsymbol{\omega}\|^2} \| = \left(\int \|\boldsymbol{\omega}\|^4 e^{-16\pi^2\sigma^2\|\boldsymbol{\omega}\|^2} d\boldsymbol{\omega} \right)^{\frac{1}{2}} \tag{7}$$

$$\propto \sigma^{\frac{-5}{2}} \tag{8}$$

Therefore, the cosine similarity for the Gaussian perturbation kernel can be computed as follows:

$$\text{Similarity}(\mathrm{S}_{\sqrt{p}}, \mathrm{S}_f) \propto \sigma^{\frac{5}{2}} \langle \mathrm{S}_f, \|\boldsymbol{\omega}\|^2 \mathrm{S}_{\sqrt{p}} \rangle$$

Using the L1 norm for the squared gradients of flattened explanations. We have observed experimentally that S_f depends on the input scale, hence we use a convex interpolation with noise to support the assumption that S_f remains constant (see Appendix E.2 and Fig. 8). The results are shown in Fig. 4.

Remark 6. The optimal perturbation scale, determined by maximizing cosine similarity, offers a standard method for selecting the perturbation scale in SG^2. Given that user discretion in the selection of perturbation scale contributes to inconsistencies, as illustrated in Fig. 1, opting for the optimal perturbation scale mitigates these inconsistencies in the results. Hence, *it represents a more suitable design choice.* In Sec. 5 we demonstrate that this approach not only enhances SG^2's robustness against inconsistencies but also its adaptability to diverse datasets, which in turn can improve the quantitative performance.

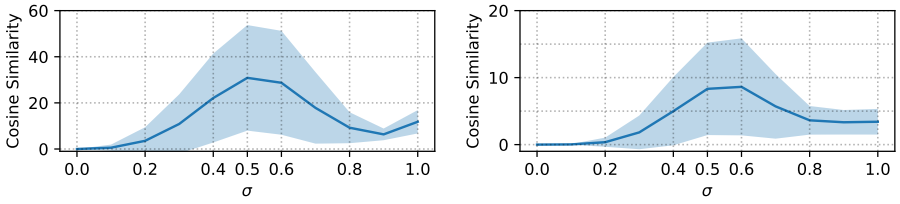
4.2 Aggregation of Information With Explanation Prior

We have shown that explanations are influenced by specific frequency bands, rather than representing the importance of individual pixels for prediction. Instead, *they indicate the contributions within a specific frequency range.* This observation motivates us to investigate a method for aggregation of information across different bands, which closely resembles an ensemble of IG and SG [55,57] which we call SpectralLens.

Definition 3. ***SpectralLens*** *is an ensemble explanation method, that aggregates information across different frequency bands as follows*

$$\mathbb{E}_{p(\sigma)} \left[\mathbb{E}_{p(\tilde{\boldsymbol{x}})} \left[(\nabla f(\tilde{\boldsymbol{x}}))^2 \right] \right] \propto \iint \mathrm{S}_f(\boldsymbol{\omega}) \|\boldsymbol{\omega}\|^2 e^{-8\pi^2\sigma\|\boldsymbol{\omega}\|^2} d\boldsymbol{\omega} p(\sigma) d\sigma \tag{9}$$

Where we refer to this method by SL^2 as it is using the squared gradients and we call $p(\sigma)$ as an explanation prior, gauging the frequency bands that we choose to represent in an explanation.



(a) (b)

Fig. 4: Cosine similarity between kernel and classifier, defined in Eq. (6), is measured for the Gaussian perturbation kernel and ResNet50 averaged on the 10K images of the ImageNet validation set (a) and Food101 (b). This metric assesses the similarity between the PSD of the perturbation kernel and that of the classifier, indicating the perturbation level's effectiveness for kernels with finite norm in the Fourier domain. When using a Dirac explanation prior in Eq. (9), where a perturbation level is assigned per image or dataset, the optimal choice for the perturbation level is where the similarity is maximized across the image or dataset respectively. Note that this level of noise refers to a band-pass filter, shown in Fig. 2, which suggests the consideration of a latent distribution on the *frequency of labeling*, reflecting the workforce's mindset during data labeling.

Remark 7. SL^2 exhibits robustness against inconsistencies in SG^2 due to its aggregation of information from various bands. Therefore, *it is a preferable design choice not to employ a Dirac explanation prior with* SG^2, as it eliminates inconsistencies, depicted in Fig. 1. Nevertheless, as demonstrated in the Sec. 5, this approach may enhance the quantitative performance of explanations.

Intuition. Considering Eq. (4), we realize that the combination of perturbation and gradient explainer creates overlapping band-pass carrier waves that modulate the amplitude of the spectral density of the classifier. Therefore, a new summary statistic is required that can reflect a wide range of frequency bands. We explicitly identified the spectral properties of the explanations, recognizing that not all frequencies are equally desired. Explanations with very high frequencies primarily attribute predictions to pixels. Therefore, we utilize an *explanation prior* $p(\sigma)$ for a nuanced control of the spectral properties of explanations. Although there are many possible alternatives for selecting an explanation prior, we used $p(\sigma) = \mathrm{Unif}(0, 1)$ for simplicity.

Definition 4. *We define* ***ArgLens*** *(AL) as the point at which the* SG^2 *is maximized for pixel* x_i*. Formally defined as*

$$\mathrm{ArgLens}(x_i) = \arg\max_{\sigma} \left\{ \mathrm{SG}^2(x_i) \right\} \tag{10}$$

where σ *is the hyperparameter of the perturbation kernel.*

AL shows the frequency band at which the pixel's attribution is maximized. Alternatively, it can be interpreted as the robustness of a pixel to noise. We

have visualized the AL for the Gaussian kernel in Fig. 5 and discussed the way it helps the interpretation of explanations in Sec. 4.2.

The aggregation of explanations from different frequencies is visualized in Fig. 5. Moreover, the basic explanation methods are obtainable with a Dirac explanation prior $p(\sigma) = \delta(\sigma)$. Therefore, it can be used as a unified formulation for gradient-based explanation methods in spectral domain.

Interesting Properties of SpectralLens and ArgLens. In the absence of a robust mathematical framework, the interpretation of explanations may be subjective. Thus, explanation methods must be grounded in solid theoretical principles. SL^2 addresses potential inconsistencies by aggregating attributions across frequency bands, ensuring fair representation of all pixels' contributions.

As SG^2 results in bands-pass filters (see Sec. 3.3), its attributions reflect the changes in the decision boundary within a certain frequency band. Furthermore, AL reveals the frequency band at which the attribution is maximized. Hence, lower AL indicates that the attributions are likely governed by high-frequency components of the decision boundary, more sensitive to input noise, consequently less robust. An ideal classifier with smooth decision boundaries, *i.e.* robust features, leads to high AL everywhere, with a high intensity in SL^2 on relevant pixels and low intensity on irrelevant pixels.

Given the different purposes of pixels within an image, the frequency of activation may vary and can be visualized via AL. Since SL^2 and AL offer complementary information, they should be interpreted together. This approach offers a deeper understanding compared to a simplistic perspective, as it has been shown by Rahaman *et al.* [43] that high-frequency regions emerge later in training, are more challenging to fit, and are prone to change. On the contrary, low-frequency components of the decision boundary manifest earlier in training, are easier to fit, and exhibit greater stability with respect to noise.

It is important to clarify that "high-frequency" refers to features considered high-frequency by the model, and not necessarily frequency across spatial dimensions (*e.g.* width, height). Nonetheless, empirical findings demonstrate that high-frequency features in spatial domain are more likely to be identified as high frequency by the model (see Appendix I).

5 Evaluation

Given the experimental discoveries that squared gradients lead to better results [30], we evaluate both scenarios for different hyperparameters and compare them with SL in Tabs. 1 and 2, refer to Appendix E for implementation details and reproducibility. For the purposes of this study, we use pixel removal strategy, despite its problems [30]. The spectral analysis of pixel removal strategy and possible solutions are deferred to future research.

We report single-pixel most-important-first (MoRF) insertion and deletion scores of the methods to assess the design choices previously proposed by our theoretical framework: (1) utilizing squared gradients instead of gradients in

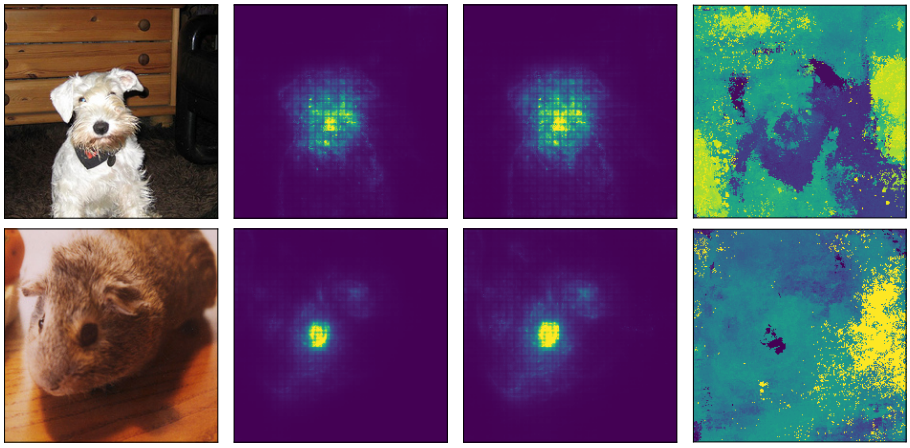


Fig. 5: This figure shows example visualizations of explanations generated by SG^2 implementing design decisions suggested in our work in Remarks 6 and 7. (b) shows the SG^2 visualized at the maximum similarity according to Fig. 4, for this image $\sigma = 0.5$. (c) shows the SL attributions. Lastly, (d) shows the frequencies obtained from AL (defined in Eq. (10)), which correspond to the perturbation level at which SG^2 is maximized. Since we are interested in frequencies rather than the actual perturbation levels, we visualize 2 stds above and below the mean of frequencies obtained by $\omega = 1/(1 + \text{AL})$ (see Appendix I for more examples). The heatmaps in (b) and (c) show how much the model is relying on the pixel, while (d) shows the frequency at which the feature's contribution maximizes. In the provided examples, we can see that the model is relying on the nose (top) and eye (bottom) while the same region is of a lower frequency compared to the head. This can be interpreted as the model finding the nose and eye more stable features compared to the head.

Remark 5, (2) finding a standard perturbation scale in Remark 6, and (3) employing a uniform explanation prior instead of Dirac, stated in Remark 7 (see Fig. 7 for an ablation study of kernel variance and Appendix G for an ablation study of patch removal instead of single pixel removal). We evaluated the suggested design choices on ImageNet [20] and Food101 [1] in Tab. 1. However, smaller datasets, such as CIFAR, are omitted because their resolution restricts the potential gap between high- and low-frequency features.

Limitations of Our Framework. The choice of independent Gaussian perturbation is limited due to its inability to incorporate prior knowledge about the input space or data manifold characteristics, and other data modalities such as graphs and time series. However, preferring simpler analytical forms within our framework limits the alternative choices. Efficiency is another challenge compared to basic methods, as performing back-props at various noise scales is costly. Moreover, our attribution of frequencies in visualizations, such as SL, may lack the desired generalizability across data points, as high-frequency re-

Table 1: This table presents the assessment of the proposed design choices in Remarks 5 to 7 across different methods, highlighting statistically significant improvements with bold fonts. We conduct t-tests comparing gradient methods to their squared-gradient counterparts to evaluate the design choice in Remark 5. Furthermore, we conduct t-tests between SG^2 with the original heuristic and those ensuring consistency to evaluate design decisions in Remarks 6 and 7. Overlines and underlines denote highest or lowest values within columns, respectively. Although there are no squared versions of Occlusion Sensitivity (OS) [5] and RISE [42], they are included as low-frequency baselines. While the dimensions of boolean masks utilized in OS and RISE can implicitly regulate the sampling variance (see Appendix C), the exponential increase in the required number of samples in high-frequency explanations renders them less practical. SG scores follow the heuristic of the SmoothGrad paper [55] *i.e.* $\sigma = (\max(\boldsymbol{x}) - \min(\boldsymbol{x}))\,0.1$. XIG indicates the input multiplication in the original implementation of IntegratedGrad [57], while IG does not apply such transformation. Rankings for methods using gradients are obtained from negative gradients, including RISE and Occlusion Sensitivity, according to Eq. (2). Furthermore, $\mathrm{SG}^2_{\mathrm{Opt.}}$ uses an optimized kernel per image and SL^2 uses a uniform explanation prior. Comparing SG^2 and $\mathrm{SG}^2_{\mathrm{Opt.}}$, we can see that $\mathrm{SG}^2_{\mathrm{Opt.}}$ can adapt to the statistics of the Food101. Therefore, $\mathrm{SG}^2_{\mathrm{Opt.}}$, while not subject to inconsistencies (shown in Fig. 1), can improve the performance due to its adaptability. Hence, we can ensure the consistency of SG^2 without compromising its performance. Also, note that the input multiplication enhances visual quality, but degrades the performance.

		ImageNet		Food101	
		Insertion(↑)	Deletion(↓)	Insertion(↑)	Deletion(↓)
Gradients	OS	0.384 ± 0.008	0.237 ± 0.010	0.378 ± 0.011	0.209 ± 0.009
	RISE	$\overline{0.494} \pm 0.014$	0.174 ± 0.007	$\overline{0.410} \pm 0.013$	0.198 ± 0.007
	SG	0.212 ± 0.010	0.179 ± 0.009	0.179 ± 0.008	$\underline{0.137} \pm 0.007$
	XIG	0.208 ± 0.009	0.204 ± 0.008	0.101 ± 0.007	0.177 ± 0.007
	IG	0.201 ± 0.010	0.182 ± 0.009	0.145 ± 0.006	$\underline{0.131} \pm 0.005$
$\mathrm{Gradients}^2$	SG^2	$\mathbf{0.465} \pm 0.013$	$\underline{\mathbf{0.157}} \pm 0.007$	$\mathbf{0.371} \pm 0.012$	0.162 ± 0.006
	XIG^2	$\mathbf{0.334} \pm 0.014$	$\mathbf{0.187} \pm 0.008$	$\mathbf{0.204} \pm 0.010$	0.263 ± 0.007
	IG^2	$\mathbf{0.340} \pm 0.012$	$\mathbf{0.168} \pm 0.007$	$\mathbf{0.320} \pm 0.009$	$\underline{0.135} \pm 0.005$
	$\mathrm{SG}^2_{\mathrm{Opt.}}$	$\overline{\mathbf{0.491}} \pm 0.015$	0.187 ± 0.007	$\mathbf{0.394} \pm 0.013$	0.167 ± 0.006
	SL^2	$\overline{\mathbf{0.493}} \pm 0.016$	0.169 ± 0.007	$\mathbf{0.395} \pm 0.014$	0.157 ± 0.007

gions in one image may correspond to low-frequency regions in another image (see Appendix I). Identifying generalizable attributions across images remains a subject for future investigation.

6 Related Works

Probabilistic and Spectral View of Explanations. Despite sporadic attempts to analyze explanations from a probabilistic view [17, 59], this view is more common in the literature around the uncertainty quantification of expla-

nations [29,40,54]. Our work is closely related to Covert *et al.* [19], if their work is seen from a probabilistic perspective. Aside from the works around the spectral view of deep networks [10, 11, 15, 43, 58], as of our current knowledge, there is no work adopting a spectral view of explanations along pixels' value. Nonetheless, there are works that build on a spectral view along spatial dimensions [34, 35], extending an optimization-based definition of explanations [22].

Unifying Views to Explanations. Covert *et al.* [19] propose an information theoretic framework that unifies several explanation methods, focusing on three design choices: (1) feature removal method, (2) which model behavior to analyze, and (3) summary statistics. Our probabilistic framework, being similar to Covert's, offers a distinct viewpoint relevant to our spectral analysis. It can be used as a unifying view of explanations (see Appendix C), inheriting from Covert's framework. Notably, our work aligns with the broader context of research under finding unifying representations of explanations [3, 26, 37].

Input-space or Function-space Perturbations. Methods such as NoiseGrad [14], Grad-CAM [51], RectifiedGrad [32], Layer-wise Relevance Propagation [7], DeepLift [53], and even Sharpness-Aware Minimization (SAM) [24] employ function space perturbations. Utilizing such distributions offers data modality independence, whereas input space perturbation remains agnostic about the model architecture. While acknowledging the duality between function and input space distributions, our focus on explicit input space distributions excludes these methods from our study. Although we recognize the utility of integrating function space perturbation into our mathematical model, we defer this aspect to future investigations. It can be noted, however, that extending this work to function space perturbation will connect it to the literature on double descent and smooth interpolation of samples [25, 41].

7 Conclusion

In this study, we deepen understanding of pixel attribution methodologies by rigorously analyzing their theoretical underpinnings, revealing a common spectral bias. Our analysis supports the experimental finding of recent works regarding the efficacy of using squared gradients. Moreover, we demonstrate how spectral properties can lead to inconsistent explanations and propose two design choices informed by our theoretical framework to address these inconsistencies: standardizing hyperparameters based on cosine similarity between the perturbation kernel and classifier, and aggregating explanations with a given prior. Despite the effectiveness of our approach, we acknowledge limitations and suggest avenues for further research. Lastly, we assess the suggested design choices using deletion and insertion scores.

Acknowledgements

We would like to thank Giovanni Luca Marchetti for an early review of our work and his kind feedback. 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 Centre at Linköping University and the Knut and Alice Wallenberg foundation.

References

1. Food-101 – Mining Discriminative Components with Random Forests, https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity Checks for Saliency Maps (Nov 2020). https://doi.org/10.48550/arXiv.1810.03292, http://arxiv.org/abs/1810.03292, arXiv:1810.03292 [cs, stat]
3. Agarwal, S., Jabbari, S., Agarwal, C., Upadhyay, S., Wu, Z.S., Lakkaraju, H.: Towards the Unification and Robustness of Perturbation and Gradient Based Explanations (Jul 2021). https://doi.org/10.48550/arXiv.2102.10618, http://arxiv.org/abs/2102.10618, arXiv:2102.10618 [cs]
4. Alvarez-Melis, D., Jaakkola, T.S.: On the Robustness of Interpretability Methods (Jun 2018). https://doi.org/10.48550/arXiv.1806.08049, http://arxiv.org/abs/1806.08049, arXiv:1806.08049 [cs, stat]
5. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for Deep Neural Networks (Mar 2018). https://doi.org/10.48550/arXiv.1711.06104, http://arxiv.org/abs/1711.06104, arXiv:1711.06104 [cs, stat]
6. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI (Dec 2019). https://doi.org/10.48550/arXiv.1910.10045, http://arxiv.org/abs/1910.10045, arXiv:1910.10045 [cs]
7. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE **10**(7), e0130140 (Jul 2015). https://doi.org/10.1371/journal.pone.0130140, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140
8. Balduzzi, D., Frean, M., Leary, L., Lewis, J.P., Ma, K.W.D., McWilliams, B.: The Shattered Gradients Problem: If resnets are the answer, then what is the question? (Jun 2018). https://doi.org/10.48550/arXiv.1702.08591, http://arxiv.org/abs/1702.08591, arXiv:1702.08591 [cs, stat]
9. Bansal, N., Agarwal, C., Nguyen, A.: SAM: The Sensitivity of Attribution Methods to Hyperparameters (Apr 2020). https://doi.org/10.48550/arXiv.2003.08754, http://arxiv.org/abs/2003.08754, arXiv:2003.08754 [cs]
10. Basri, R., Galun, M., Geifman, A., Jacobs, D., Kasten, Y., Kritchman, S.: Frequency Bias in Neural Networks for Input of Non-Uniform Density (Mar 2020), https://api.semanticscholar.org/CorpusID:212644664

11. Benbarka, N., Hofer, T., Ul-Moqeet Riaz, H., Zell, A.: Seeing Implicit Neural Representations as Fourier Series. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 2283–2292 (Jan 2022). https://doi.org/10.1109/WACV51458.2022.00234, https://ieeexplore.ieee.org/document/9706795/
12. Blackman, R.B., Tukey, J.W.: The measurement of power spectra from the point of view of communications engineering — Part I. The Bell System Technical Journal **37**(1), 185–282 (Jan 1958). https://doi.org/10.1002/j.1538-7305.1958.tb03874.x, http://ieeexplore.ieee.org/document/6768513/
13. Brughmans, D., Melis, L., Martens, D.: Disagreement amongst counterfactual explanations: How transparency can be deceptive (Apr 2023). https://doi.org/10.48550/arXiv.2304.12667, http://arxiv.org/abs/2304.12667, arXiv:2304.12667 [cs]
14. Bykov, K., Hedström, A., Nakajima, S., Höhne, M.M.C.: NoiseGrad: Enhancing Explanations by Introducing Stochasticity to Model Weights (May 2022). https://doi.org/10.48550/arXiv.2106.10185, http://arxiv.org/abs/2106.10185, arXiv:2106.10185 [cs]
15. Cao, Y., Fang, Z., Wu, Y., Zhou, D.X., Gu, Q.: Towards Understanding the Spectral Bias of Deep Learning. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. pp. 2205–2211. International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada (Aug 2021). https://doi.org/10.24963/ijcai.2021/304, https://www.ijcai.org/proceedings/2021/304
16. Chen, H., Janizek, J.D., Lundberg, S., Lee, S.I.: True to the Model or True to the Data? (Jun 2020). https://doi.org/10.48550/arXiv.2006.16234, http://arxiv.org/abs/2006.16234, arXiv:2006.16234 [cs, stat]
17. Chen, J., Song, L., Wainwright, M.J., Jordan, M.I.: Learning to Explain: An Information-Theoretic Perspective on Model Interpretation (Jun 2018). https://doi.org/10.48550/arXiv.1802.07814, http://arxiv.org/abs/1802.07814, arXiv:1802.07814 [cs, stat]
18. Cohen, L.: The generalization of the Wiener-Khinchin theorem. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181). vol. 3, pp. 1577–1580. IEEE (1998), https://ieeexplore.ieee.org/abstract/document/681753/
19. Covert, I., Lundberg, S., Lee, S.I.: Explaining by Removing: A Unified Framework for Model Explanation (May 2022). https://doi.org/10.48550/arXiv.2011.14878, http://arxiv.org/abs/2011.14878, arXiv:2011.14878 [cs, stat]
20. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (Jun 2009). https://doi.org/10.1109/CVPR.2009.5206848, https://ieeexplore.ieee.org/document/5206848, iSSN: 1063-6919
21. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Jun 2021). https://doi.org/10.48550/arXiv.2010.11929, http://arxiv.org/abs/2010.11929, arXiv:2010.11929 [cs]
22. Fong, R., Patrick, M., Vedaldi, A.: Understanding Deep Networks via Extremal Perturbations and Smooth Masks (Oct 2019). https://doi.org/10.48550/arXiv.1910.08485, http://arxiv.org/abs/1910.08485, arXiv:1910.08485 [cs, stat]

23. Fong, R., Vedaldi, A.: Interpretable Explanations of Black Boxes by Meaningful Perturbation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3449–3457 (Oct 2017). https://doi.org/10.1109/ICCV.2017.371, http://arxiv.org/abs/1704.03296, arXiv:1704.03296 [cs, stat]
24. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-Aware Minimization for Efficiently Improving Generalization (Apr 2021). https://doi.org/10.48550/arXiv.2010.01412, http://arxiv.org/abs/2010.01412, arXiv:2010.01412 [cs, stat]
25. Gamba, M., Azizpour, H., Björkman, M.: On the Lipschitz Constant of Deep Networks and Double Descent (Nov 2023). https://doi.org/10.48550/arXiv.2301.12309, http://arxiv.org/abs/2301.12309, arXiv:2301.12309 [cs]
26. Han, T., Srinivas, S., Lakkaraju, H.: Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post Hoc Explanations (Dec 2022). https://doi.org/10.48550/arXiv.2206.01254, http://arxiv.org/abs/2206.01254, arXiv:2206.01254 [cs]
27. Harel, N., Gilad-Bachrach, R., Obolski, U.: Inherent Inconsistencies of Feature Importance (Jun 2022). https://doi.org/10.48550/arXiv.2206.08204, http://arxiv.org/abs/2206.08204, arXiv:2206.08204 [cs]
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (Dec 2015). https://doi.org/10.48550/arXiv.1512.03385, http://arxiv.org/abs/1512.03385, arXiv:1512.03385 [cs]
29. Hill, D., Masoomi, A., Ghimire, S., Torop, M., Dy, J.: Explanation Uncertainty with Decision Boundary Awareness (Nov 2022). https://doi.org/10.48550/arXiv.2210.02419, http://arxiv.org/abs/2210.02419, arXiv:2210.02419 [cs]
30. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A Benchmark for Interpretability Methods in Deep Neural Networks (Nov 2019). https://doi.org/10.48550/arXiv.1806.10758, http://arxiv.org/abs/1806.10758, arXiv:1806.10758 [cs, stat]
31. Hou, X., Zhang, L.: Saliency Detection: A Spectral Residual Approach. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (Jun 2007). https://doi.org/10.1109/CVPR.2007.383267, https://ieeexplore.ieee.org/document/4270292, iSSN: 1063-6919
32. Kim, B., Seo, J., Jeon, S., Koo, J., Choe, J., Jeon, T.: Why are Saliency Maps Noisy? Cause of and Solution to Noisy Saliency Maps (Sep 2019). https://doi.org/10.48550/arXiv.1902.04893, http://arxiv.org/abs/1902.04893, arXiv:1902.04893 [cs, stat]
33. Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (Un)reliability of saliency methods (Nov 2017). https://doi.org/10.48550/arXiv.1711.00867, http://arxiv.org/abs/1711.00867, arXiv:1711.00867 [cs, stat]
34. Kolek, S., Nguyen, D.A., Levie, R., Bruna, J., Kutyniok, G.: Cartoon Explanations of Image Classifiers (Oct 2022). https://doi.org/10.48550/arXiv.2110.03485, http://arxiv.org/abs/2110.03485, arXiv:2110.03485 [cs]
35. Kolek, S., Windesheim, R., Loarca, H.A., Kutyniok, G., Levie, R.: Explaining Image Classifiers with Multiscale Directional Image Representation (Apr 2023). https://doi.org/10.48550/arXiv.2211.12857, http://arxiv.org/abs/2211.12857, arXiv:2211.12857 [cs]
36. Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., Lakkaraju, H.: The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective (Feb 2022). https://doi.org/10.48550/arXiv.2202.01602, http://arxiv.org/abs/2202.01602, arXiv:2202.01602 [cs]

37. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
38. Marmarelis, P.Z., Marmarelis, V.Z.: The White-Noise Method in System Identification. In: Marmarelis, P.Z., Marmarelis, V.Z. (eds.) Analysis of Physiological Systems: The White-Noise Approach, pp. 131–180. Computers in Biology and Medicine, Springer US, Boston, MA (1978). https://doi.org/10.1007/978-1-4613-3970-0_4, https://doi.org/10.1007/978-1-4613-3970-0_4
39. Marques-Silva, J., Ignatiev, A.: Delivering Trustworthy AI through Formal XAI. Proceedings of the AAAI Conference on Artificial Intelligence **36**(11), 12342–12350 (Jun 2022). https://doi.org/10.1609/aaai.v36i11.21499, https://ojs.aaai.org/index.php/AAAI/article/view/21499
40. Marx, C., Park, Y., Hasson, H., Wang, Y., Ermon, S., Huan, L.: But Are You Sure? An Uncertainty-Aware Perspective on Explainable AI. In: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. pp. 7375–7391. PMLR (Apr 2023), https://proceedings.mlr.press/v206/marx23a.html
41. Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., Sutskever, I.: Deep Double Descent: Where Bigger Models and More Data Hurt (Dec 2019). https://doi.org/10.48550/arXiv.1912.02292, http://arxiv.org/abs/1912.02292, arXiv:1912.02292 [cs, stat]
42. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models (Sep 2018). https://doi.org/10.48550/arXiv.1806.07421, http://arxiv.org/abs/1806.07421, arXiv:1806.07421 [cs]
43. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F.A., Bengio, Y., Courville, A.: On the Spectral Bias of Neural Networks (May 2019). https://doi.org/10.48550/arXiv.1806.08734, http://arxiv.org/abs/1806.08734, arXiv:1806.08734 [cs, stat]
44. Rahman, M.: Applications of Fourier Transforms to Generalized Functions. WIT Press (2011), google-Books-ID: k_rdcKaUdr4C
45. Ras, G., van Gerven, M., Haselager, P.: Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. In: Escalante, H.J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., van Gerven, M. (eds.) Explainable and Interpretable Models in Computer Vision and Machine Learning, pp. 19–36. The Springer Series on Challenges in Machine Learning, Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-98131-4_2, https://doi.org/10.1007/978-3-319-98131-4_2
46. Reiter, H., Stegeman, J.D.: Classical Harmonic Analysis and Locally Compact Groups. Oxford University Press (Nov 2000). https://doi.org/10.1093/oso/9780198511892.001.0001, https://academic.oup.com/book/54460
47. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier (Aug 2016). https://doi.org/10.48550/arXiv.1602.04938, http://arxiv.org/abs/1602.04938, arXiv:1602.04938 [cs, stat]
48. Rong, Y., Leemann, T., Borisov, V., Kasneci, G., Kasneci, E.: A Consistent and Efficient Evaluation Strategy for Attribution Methods (Jun 2022). https://doi.org/10.48550/arXiv.2202.00449, http://arxiv.org/abs/2202.00449, arXiv:2202.00449 [cs]
49. 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 (May 2019). https://doi.org/10.1038/s42256-019-0048-x, https://www.nature.com/articles/s42256-019-0048-x
50. Rudin, W.: Principles of Mathematical Analysis. McGraw-Hill (1976), google-Books-ID: kwqzPAAACAAJ

51. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision **128**(2), 336–359 (Feb 2020). https://doi.org/10.1007/s11263-019-01228-7, http://arxiv.org/abs/1610.02391, arXiv:1610.02391 [cs]
52. Serov, V.: The Riemann–Lebesgue Lemma. In: Serov, V. (ed.) Fourier Series, Fourier Transform and Their Applications to Mathematical Physics, pp. 33–35. Applied Mathematical Sciences, Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-65262-7_6, https://doi.org/10.1007/978-3-319-65262-7_6
53. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not Just a Black Box: Learning Important Features Through Propagating Activation Differences (Apr 2017). https://doi.org/10.48550/arXiv.1605.01713, http://arxiv.org/abs/1605.01713, arXiv:1605.01713 [cs]
54. Slack, D., Hilgard, S., Singh, S., Lakkaraju, H.: Reliable Post hoc Explanations: Modeling Uncertainty in Explainability (Nov 2021). https://doi.org/10.48550/arXiv.2008.05030, http://arxiv.org/abs/2008.05030, arXiv:2008.05030 [cs, stat]
55. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise (Jun 2017). https://doi.org/10.48550/arXiv.1706.03825, http://arxiv.org/abs/1706.03825, arXiv:1706.03825 [cs, stat]
56. Stoica, P., Moses, R.L.: Spectral analysis of signals, vol. 452. Pearson Prentice Hall Upper Saddle River, NJ (2005), http://user.it.uu.se/~ps/SAS-new.pdf
57. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks (Jun 2017). https://doi.org/10.48550/arXiv.1703.01365, http://arxiv.org/abs/1703.01365, arXiv:1703.01365 [cs]
58. Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. ArXiv (Jun 2020), https://api.semanticscholar.org/CorpusID:219791950
59. Wang, E., Khosravi, P., Broeck, G.V.d.: Probabilistic Sufficient Explanations (May 2021). https://doi.org/10.48550/arXiv.2105.10118, http://arxiv.org/abs/2105.10118, arXiv:2105.10118 [cs]
60. Wang, Y., Wang, X.: "Why Not Other Classes?": Towards Class-Contrastive Back-Propagation Explanations. Advances in Neural Information Processing Systems **35**, 9085–9097 (Dec 2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/3b7a66b2d1258e892c89f485b8f896e0-Abstract-Conference.html
61. Wilming, R., Kieslich, L., Clark, B., Haufe, S.: Theoretical Behavior of XAI Methods in the Presence of Suppressor Variables (Jun 2023). https://doi.org/10.48550/arXiv.2306.01464, http://arxiv.org/abs/2306.01464, arXiv:2306.01464 [cs, stat]
62. Zeiler, M.D., Fergus, R.: Visualizing and Understanding Convolutional Networks (Nov 2013). https://doi.org/10.48550/arXiv.1311.2901, http://arxiv.org/abs/1311.2901, arXiv:1311.2901 [cs]