A. Explanation Methods are Low-pass Filters

This section builds on key insights from studies focused on the spectral analysis of gradient-based explanation methods, which provides a foundation for understanding how these methods interact with different frequency components.

Gradient-based explanation methods often [1, 24] incorporate a perturbation mechanism to reduce noise in raw gradients, as seen in VanillaGrad. This perturbation can be represented as a probability distribution $p(x)$ in the input space of the neural network. For mathematical clarity, we focus on raw gradients, though the spectral properties of squared gradients are also covered in [41].

The process of sampling and averaging can be formulated as an expectation over the perturbation distribution:

$$\mathbb{E}_{\mathcal{N}(x)}[\nabla f(x)] \tag{9}$$

Using this formulation, we can derive a spectral representation of explanation methods:

$$\mathbb{E}_{\mathcal{N}(x)}[\nabla f(x)] \propto \int \omega \widehat{f}(\omega) \widehat{\mathcal{N}}(\omega) d\omega \tag{10}$$

where $\widehat{f}(\omega)$ denotes the Fourier transform of the neural network, $\widehat{\mathcal{N}}(\omega)$ the Fourier transform of the perturbation distribution, and ω arises as a scaling factor due to the Fourier transform of the gradient.

This equation highlights the inherent filtering behavior of gradient-based explanation methods. The gradient operator acts as a high-pass filter, emphasizing high-frequency components of the model function, while the perturbation mechanism (*e.g.*, Gaussian noise in SmoothGrad) serves as a low-pass filter, attenuating high-frequency components. The combined effect forms a band-pass filter, which selectively attributes importance to features within specific frequency ranges. This interplay between gradient computation and input perturbation fundamentally shapes the behavior of gradient-based explanations.

Given this behavior, we focus solely on the VanillaGrad in this work and disregard the variations in the neighborhoods, *i.e.* assuming $\mathcal{N}(x) = \delta(x)$, in our theoretical analysis.

B. An Empirical Study of Sharpness via Tail

In this work, we have primarily examined the impact of ReLU on the tail of the network's power spectrum in the main text, as it is a prevalent choice in many architectures used in contemporary computer vision. However, on the empirical side, our approach is not limited to this activation function, and similar analyses can be extended to other architectural choices. In this section, we empirically investigate the effects of other design decisions on the tail of the power spectrum.

To isolate the effect of the smooth parameterization of ReLU, we applied a validation accuracy cap to minimize the influence of initialization on our results. Since ReLU networks leverage well-established initialization schemes and generally achieve better convergence than our smooth parameterization, setting a high accuracy cap could introduce initialization as a confounding variable in our analysis. This approach allows for a meaningful comparison by ensuring networks are evaluated under similar training budgets.

To investigate the impact of this validation accuracy cap, we conduct an ablation study where we remove the cap and train various networks with smooth ReLU parameterizations for approximately 200 epochs, continuing until their learning curves plateau.

The spatial power spectra for runs with different learning rates are shown in Fig. 6. Under these conditions, we can analyze the effect of the smoothness parameter β on both training and test accuracy–see Fig. 5.

Continuing our ablation study with learning rate and depth, Fig. 7 shows that the tail of the spatial power spectrum in a network with lower smoothness parameter β contains less high-frequency content compared to a network with standard ReLU activations. This aligns well with earlier findings on NTK [7], regarding the invariance of NTK with respect to depth.

As previously discussed, input size plays a crucial role in our experiments and has been examined in prior studies by varying the dataset. To isolate the effect of input size alone, we trained models on different versions of the Imagenette dataset with input sizes of 224×224, 122×122, 64×64, and 46×46. The results are presented in Fig. 8.

We have also examined the impact of skip connections and batch normalization on the tail of the spatial power spectrum, results shown in Fig. 9. To be able to include skip connections, we have used slightly deeper networks. While skip connections slightly affect the tail, batch normalization generally amplifies it significantly. Interestingly, this agrees with [12] on the effect of batch normalization on learning high frequency information, yet suggesting a potential research direction to reconcile this observation with prior findings on skip connections mitigating gradient noise [4] In all cases, ReLU contributes to a heavier tail, whereas smoother versions reduce the expected frequency as defined in Eq. (1).

We investigated the impact of input noise during training on the tail of the spatial power spectrum. Our observations indicate that Gaussian isotropic noise has a negligible effect, so we omitted the results to avoid redundancy.

C. A Short Introduction to Kernel Methods

A distinct line of research in machine learning aims to connect neural networks with the classical framework of kernel methods. This work introduced Neural Tangent Ker-

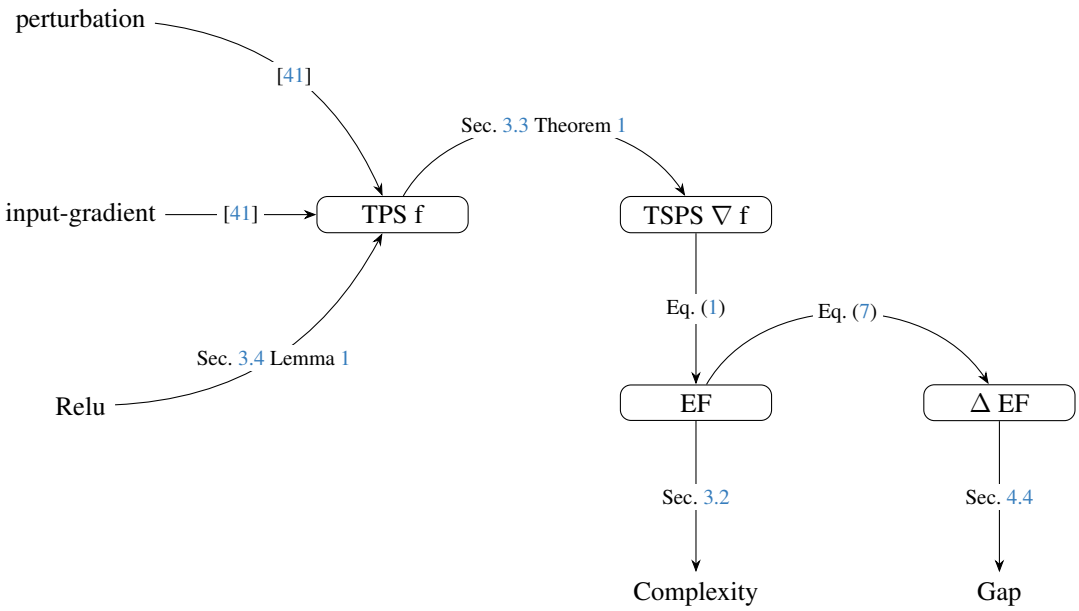


Figure 4. **Overview of Key Theoretical Connections.** A graphical overview of our contributions, which is based on prior works on spectral analysis of gradient-based explanation methods. The diagram illustrates the conceptual flow of the narrative presented in this paper. From TPS f to EF and Δ EF, highlighting key theoretical results and their related sections for analyzing complexity and explanation gap in a unified theoretical framework. A follow-up for this work is finding out how other architectural components affect TSP of a network.

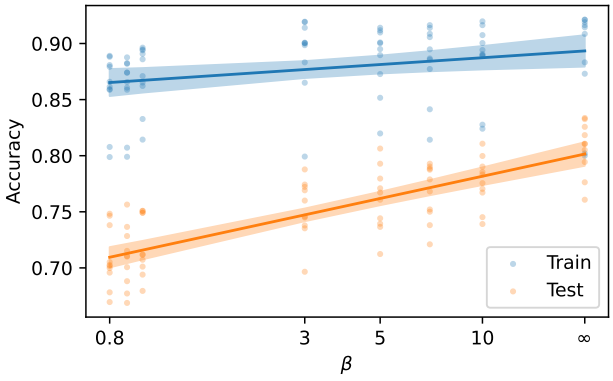
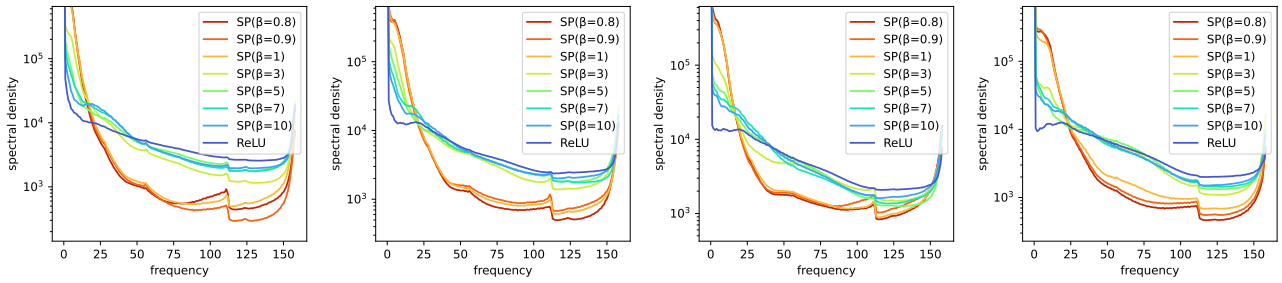


Figure 5. **Ablation Study: Impact of Smoothness Parameter on Validation Accuracy.** This figure presents an ablation study of our decision to impose an accuracy cap as an early stopping mechanism on Imagenette (224×224). By relaxing this constraint, we train smooth parameterizations of ReLU networks with varying β parameters (shown on the x-axis), where $\beta \to \infty$ corresponds to a standard ReLU network. As expected from the complexity-explainability tradeoff, restricting the network's ability to learn high-frequency information results in a lower validation accuracy.

Dataset	Input Size	LR	Depth	Cap
Imagenette	224×224	$1e-4$	5	60%
Imagenette	112×112	$1e-4$	5	60%
Imagenette	64×64	$3e-4$	5	60%
Imagenette	46×46	$5e-4$	5	60%
CIFAR10	32×32	$3e-3$	4	70%
Fashion MNIST	28×28	$1e-4$	3	80%

Table 3. **Table of hyperparameters.** This table outlines the general hyperparameters used in our experiments analyzing the tail behavior of the power spectrum (TSPS) of gradient and its relation to that of the tail of the power spectrum (TPS) of the network. LR represents the learning rate, and Cap refers to an early stopping criterion based on validation accuracy. To isolate the effect of smooth parameterization of ReLU, we implemented the validation accuracy cap to reduce the impact of initialization on our findings. Since ReLU networks benefit from well-established initialization strategies and tend to exhibit better convergence properties compared to its smooth parameterization, a high accuracy cap could introduce initialization as a confounding factor to our analysis, see Appendix B for an ablation study of this decision. This strategy ensures a meaningful comparison by comparing networks at similar training budgets.

nels (NTK) [28], which help explain certain behaviors observed during neural network training. To ensure clarity, we briefly revisit the definition of kernels: a kernel is a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that the matrix $K_{ij} = k(x_i, x_j)$ is positive semidefinite for an arbitrary set $\mathcal{X} = \{x_1, \ldots, x_n\}$. A kernel generally serves as a measure of similarity between two entities, such as inputs x_i and x_j.



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

Figure 6. **Ablation Study: Impact of Validation Accuracy Cap and Learning Rate on the Spatial Power Spectrum.** This figure presents an ablation study on our decision to impose a validation accuracy cap as an early stopping mechanism on Imagenette (224×224) across different learning rates. To assess the impact of this choice, we train the networks for extended sessions of approximately 200 epochs without the accuracy cap. Important to note that, in this setting, the spatial power spectra of the functions do not correspond to networks with comparable functional behavior, as their performance levels differ–see Fig. 5. This discrepancy was the primary motivation for enforcing the accuracy cap. Nevertheless, we observe that networks with higher smoothness parameter β, exhibit heavier tails in their spatial power spectra, indicating a greater tendency to learn higher-frequency information. However, this trend is less clear compared to the observations in Fig. 2.

	Method	$\downarrow$ EF + $\downarrow \Delta$ EF	
		ReLU-ViT	GELU-ViT
Imagenette	VanillaGrad [53]	$.239 + \Delta.000$	$.248 + \Delta.000$
	SmoothGrad [54]	$.239 + \Delta.000$	$.248 + \Delta.000$
	IntGrad [58]	$.244 + \Delta.005$	$.253 + \Delta.005$
	GuidedBp [55]	$.253 + \Delta.000$	$.247 + \Delta.014$
	DeepLift [51]	$.245 + \Delta.007$	$.254 + \Delta.006$
	GradCAM [50]	$.205 + \Delta.051$	$.197 + \Delta.033$
	LRP [9]	Undefined	Undefined

Table 4. **ReLU vs GELU in ViT-B16 [63].** This table reports the expected frequency (EF) from Eq. (1) and the explanation gap (ΔEF) from Eq. (7) across various post-hoc explanation methods for a ViT-B16 model trained from scratch on Imagenette, using different activation functions. All values are scaled by 10^4. The results indicate that the ViT architecture has a greater influence on lowering EF than the choice of activation function. Interestingly, the relative ordering of EF complexity between ReLU and GELU is inverted compared to theoretical expectations, which predict higher EF for ReLU. This discrepancy may stem from the fact that ViT induces a different kernel geometry [62] than the Laplace kernel assumed in our analysis [20]. Nonetheless, the activation function may still influence the smoothness (EF) and complexity (ΔEF) of explanations.

NTK provides a kernel-based perspective on neural networks, where similarity is defined in terms of weight gradients. Since our work in explainability relies on input gradients, we are particularly interested in the properties that NTK can reveal in this context. However, it is important to acknowledge that NTK was not originally designed for explainability, and purely theoretical predictions may be inaccurate. Therefore, experimental validation is crucial.

While a specialized version of NTK exists for convolutional networks—namely, the Convolutional Neural Tangent Kernel (CNTK)—we opted for a more general case, *i.e.* NTK framework to support our observations. This choice is motivated by the fact that the reasoning in Lemma 1 relies on a core component of NTK, the τ-transform. As demonstrated in our experiments, NTK provides a sufficiently accurate approximation for predicting the tail behavior of the power spectrum.

The discovery of NTKs represents a significant advancement in understanding neural networks. However, in this work, we primarily leverage results related to the sharpness of the kernel. Prior studies [20] have identified a striking similarity between NTKs and the Laplace kernel, under certain technical conditions, see [20]. Given that we are only interested in the tail behavior of the power spectrum, without loss of generality, we replace NTK with the Laplace kernel to simplify our analysis.

Our focus on the spectral decay properties of kernels induced by ReLU networks connects our work to studies on Reproducing Kernel Hilbert Spaces (RKHS), particularly NTK, the pre-activation tangent kernel (PTK), and related research [5, 8, 17, 20, 52] (see Appendices B and D in [5]). These connections suggest that insights in one domain may inform advancements in the other.

Additionally, our work is related to feature selection using kernels [2, 15, 21, 27], highlighting the nuanced relationship between feature selection and explainability.

This section is based on key results from the kernel methods literature, particularly [30], which serves as a valuable resource for a deeper exploration. Kernel methods provide a powerful framework for non-parametric learning by implicitly mapping data into a high-dimensional feature space

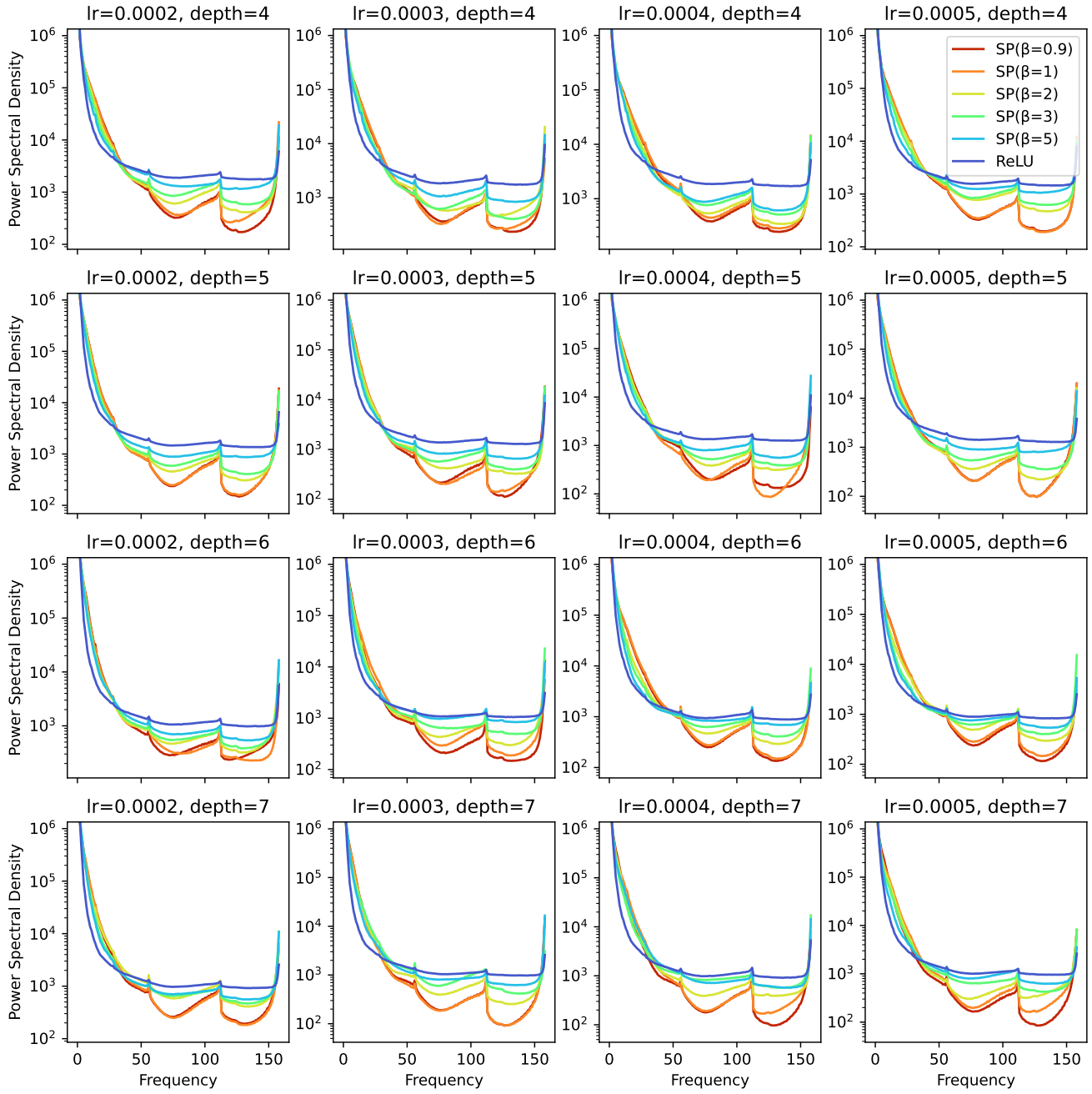
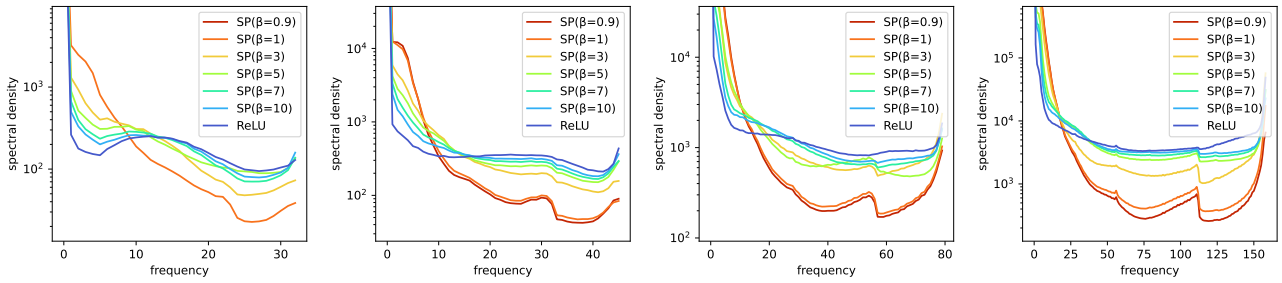


Figure 7. **Ablation Study: Impact of Depth and Learning Rate on the Spatial Power Spectrum.** This figure illustrates an ablation study on network depth (varied across rows) and learning rate (varied across columns) to assess their influence on the spatial power spectrum of the explanations. Empirically, the spatial power spectra remain largely unchanged with depth, aligning with earlier theoretical findings on NTK [7]. Additionally, the results support our theoretical framework, which predicts that increasing the smoothness parameter β, leads to heavier tails and, consequently, more complex explanations. We should note that the ReLU activation function is recovered whith $\beta \to \infty$, and all experiments were conducted on Imagenette 224×224.

through a kernel function $k(x, x')$.

Definition 1. *A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, is called a positive semidefinite kernel, if the matrix $K_{ij} = k(x_i, x_j)$ is positive semidefinite for an arbitrary non-empty set $\mathcal{X} = \{x_1, \ldots, x_n\}$.*



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

Figure 8. **Ablation Study: Impact of Input Size on the Spatial Power Spectrum.** This figure presents an ablation study, isolating the effect of input size by training multiple models on different versions of the Imagenette dataset with varying input resolutions. The input sizes range from **(a)** 46×46, **(b)** 64×64, **(c)** 112×112, and **(d)** 224×224. As can be observed, the overall trend remains consistent with previous findings in Fig. 2, though the peak in mid-range frequencies becomes less pronounced as the input size decreases. This corroborates our conjecture about the correspondence of mid-frequency and very high frequency peaks to the reliance of the model on edges.

Common examples of such kernels include exponential functions of the form: $k(x, x') = \exp(-|x - x'|^{\gamma})$, where the function is referred to as the Laplace kernel for $\gamma = 1$ and the Gaussian kernel for $\gamma = 2$.

A kernel, together with an associated inner product between functions, defines a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$, where functions inherit smoothness properties dictated by the choice of k. Notably, different kernels define RKHSs with varying smoothness constraints, and a key relationship between them is

$$\mathcal{H}_{\text{Gaussian}} \subset \mathcal{H}_{\text{Laplace}}. \tag{11}$$

This inclusion indicates that Gaussian RKHSs consist of smoother functions compared to those in the Laplace RKHS.

A fundamental result in kernel methods is the Representer theorem, which ensures that solutions to many learning problems can be expressed as kernel expansions:

$$f(x) = \sum_{i \in \mathcal{I}} \alpha_i k(x, x_i), \tag{12}$$

where x_i are training examples from the training set $\mathcal{X}$ indexed by $\mathcal{I}$, and α_i are learned coefficients.

A particularly important class of kernels, known as shift-invariant kernels, depends only on the absolute distance between inputs, *i.e.*, $\Delta = \|x - x'\|$. With a slight abuse of notation, such kernels can be written as $k(\Delta)$. This property allows one to take the Fourier transform of the kernel with respect to Δ, leading to a simplified yet insightful characterization of the RKHS:

$$\mathcal{H}_k = \left\{ f : \int \frac{|\mathcal{F}\{f\}|^2}{\mathcal{F}\{k\}} d\omega < \infty \right\}. \tag{13}$$

This expression relates the power spectrum of a function f to the Fourier transform of the reproducing kernel k. Intuitively, a function f belongs to the RKHS of k if the decay rate of its power spectrum is at least as fast as that of $\mathcal{F}\{k\}$, thereby constraining the sharpness of functions representable by the kernel.

Recent developments in the kernel-based understanding of deep networks have led to the discovery of the Neural Tangent Kernel (NTK) [8], which characterizes network behavior during training.

The (empirical) NTK for a network f with parameters W^{ℓ} at layer ℓ and two points x_0 and z_0 defined as follows,

$$\hat{k}_{\ell}(x_0, z_0) = \left\langle \frac{\partial f(x_0)}{\partial W^{(\ell)}}, \frac{\partial f(z_0)}{\partial W^{(\ell)}} \right\rangle \tag{14}$$

which is connected to pre-activation tangent kernel (PTK) $\mathcal{K}^{(\ell)}$ by noting that $\frac{\partial f(x_0)}{\partial W^{(\ell)}} = \frac{\partial f(x_0)}{\partial h_{\ell}} x_0^{\top}$. Therefore, we can write,

$$\hat{k}_{\ell}(x_0, z_0) = \mathcal{K}^{(\ell)}(x_0, z_0) \cdot x_{\ell}^{\top} z_{\ell}. \tag{15}$$

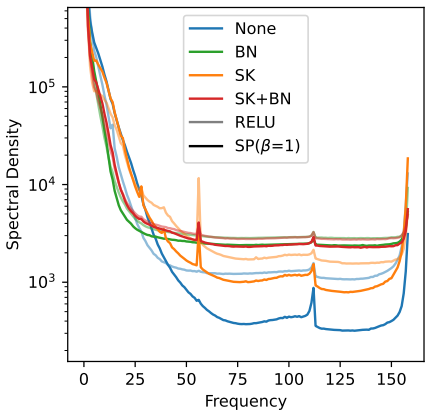
see Appendix B and D of [5] for details. Finally, since gradient-based explanation methods rely heavily on the network's input gradient, it is unsurprising that advancements in one domain can inform the other.

Notably, empirical and theoretical findings suggest that the NTK closely resembles the Laplace kernel [20], implying that the function space of neural networks is constrained similarly. This insight provides a theoretical foundation for understanding the sharpness and expressivity of neural networks through a kernel lens.

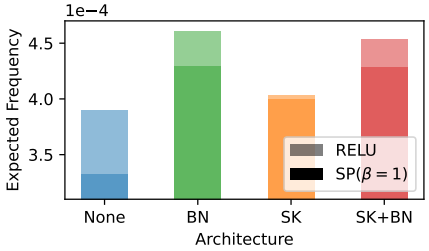
Interestingly, despite being developed for different purposes, both NTK and gradient-based explanations rely on input gradients, suggesting that insights from one field can contribute to advancements in the other.

D. CDF based Normalization

Our work is tangentially related to research on explainability, influenced by game-theoretic approaches to explana-



(a)



(b)

Figure 9. **Ablation Study: Effect of Skip Connections and Batch Normalization on the Spatial Power Spectrum Tail.** This figure illustrates the impact of skip connections and batch normalization on the tail of the spatial power spectrum on Imagenette (224×224). In **(a)** the power spectrum's tail exhibits a slight increase in networks with skip connections. In contrast, batch normalization generally contributes to a heavier tail. Nonetheless, in all cases, ReLU tends to amplify the tail, whereas replacing it with a smoothed function with parameter $\beta = 1$, reduces the expected frequency shown in **(b)**, as defined in Eq. (1).

tion [37, 39, 61]. We adopt the assumption that explanations can be represented as a ranking of input features.

While the explainability community generally agrees that explanations can be expressed as rankings, the process of obtaining these rankings remains unclear. To derive rankings, existing pipelines incorporate various normalization strategies, ranging from simple techniques such as min-max normalization to more complex, heavily engineered approaches that are harder to reproduce.

Inspired by literature on spectral analysis of signals [22, 56], we assume that an explanation method gives rise to a distribution per image across pixels. More formally, if $x(i)$ denotes the random variable observed in each pixel of explanation, which is distributed according to $\pi(x(i))$, with $\Pi(x(i))$ being its corresponding cumulative distribution function. To normalize the explanations, we use $\Pi(x(i))$ instead of the actual observed values, *i.e.* $x(i)$.

To normalize this distribution within a comparable framework, we apply the inverse transformation method. This normalization technique aligns different distributions while preserving their spectral properties and remaining insensitive to magnitude, see Fig. 13.

Compared to alternative normalization methods, such as norm or max normalization, our approach reveals clearer spectral-domain trends and is easier to reproduce, as it does not rely on extensive engineering.

While the inverse transformation method is useful in our setting, it is highly sensitive to small variations in the gradient. To mitigate this effect, we average the spectral densities over 1K samples, although meaningful results often emerge from a single image.

E. Proofs and Technical Considerations

In this section, we emphasize that the tail behavior is a relatively stable property, meaning it does not change easily. While the proof involves a tedious case analysis to rule out various edge cases, it remains conceptually straightforward.

Let k be a kernel equivalent to the Neural Tangent Kernel (NTK) of a network, and $\mathcal{I}$ is an index for our training set $\mathcal{X}$. We express our function in terms of its features as

$$f(x) = \sum_{i \in \mathcal{I}} \alpha_i k(x, x_i)$$

where, same as Eq. (12), $\mathcal{X}$ represents the training set. Since, in explainability, we are interested in the spectral decay of the function's input-gradient, we consider

$$\nabla_x f(x) = \sum_{i \in \mathcal{I}} \alpha_i \nabla_x k(x, x_i), \tag{16}$$

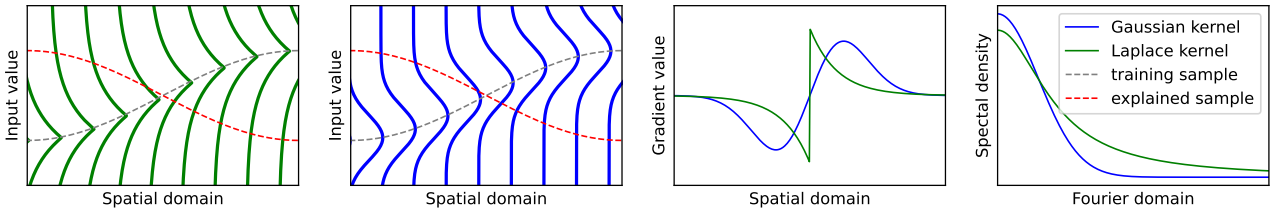
however, we may drop the subscript x from ∇_x as we only take the input-gradient.

Lemma 2. *Let $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ be a dataset of size n, and let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a shift-invariant kernel with Fourier transform $\hat{k}(\omega)$. Define the input-gradient of the kernel as $\nabla_x k(x, x')$. Then, the spectral decay of the input-gradient of k satisfies the bound:*

$$|\mathcal{F}\{\nabla_x k\}|^2 = \mathcal{O}(n\,\omega^2 |\hat{k}(\omega)|^2),$$

where $\mathcal{F}$ denotes the Fourier transform, aligned with the direction of the gradient, and ω is the frequency variable in the spectral domain.

Proof. To establish an upper bound on the tail behavior of



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

Figure 10. **Illustration of Kernel Sharpness on the Spatial Power Spectrum Tail.** A cartoon illustration depicts the theory regarding the effect of sharpness of the learned features on the tail of the power spectrum. **(a)** and **(b)**: The x-axis represents a simplified version of spatial dimensions (*e.g.*, width or height), while the y-axis shows pixel values, consequently, each sample is visualized as a line. Two samples are used for simplicity: the red line represents the sample to be explained, and the gray line represents a learned feature from the training data. In (a), the features arise from the Laplace kernel (sharp), while in column (b), they arise from the Gaussian kernel (smooth). **(c)**: By taking the gradient of the classifier with respect to the input (along the y-axis in (a)) for the red line (the sample being explained), we get a function in the spatial dimension, which is visualized in (c). **(d)**: Applying the Fourier transform to the gradient values along the spatial dimension (x-axis of (c)) reveals different decay rates for the gradient functions, which are visualized in (d). This visualization highlights that the spectral properties of the gradient values for the sample being explained depend on the spectral properties of the kernel, as formalized in Theorem 1. This visualization assumes high spatial autocorrelation between the learned features and the input, a characteristic typical of image data. For discussions on outcomes when this assumption is relaxed, refer to Appendix G. While it is well established that the NTK's spectral properties closely resemble those of the Laplace kernel, here we use the Gaussian kernel purely as an illustrative example and do not explicitly characterize the kernel corresponding to a smoother variant of ReLU.

the gradient of the kernel, we write

$$|\mathcal{F}\{\nabla f(x)\}|^2 = \left|\mathcal{F}\left\{\sum_{x_i\in\mathcal{I}} \alpha_i \nabla k(x, x_i)\right\}\right|^2 \tag{17}$$

$$\propto \omega^2 \left|\sum_{i\in\mathcal{I}} \alpha_i \mathcal{F}\{k(x, x_i)\}\right|^2 \tag{18}$$

$$\leq \omega^2 \sum_{i\in\mathcal{I}} \alpha_i^2 \left|\mathcal{F}\{k(x, x_i)\}\right|^2 \tag{19}$$

$$\leq \omega^2 |\widehat{k}(\omega)|^2 \sum_{i\in\mathcal{I}} \alpha_i^2 \tag{20}$$

$$\in \mathcal{O}\left(n\,\omega^2 |\widehat{k}(\omega)|^2\right) \tag{21}$$

□

Remark 4. *As can be seen, the bound* $\mathcal{O}\left(n\omega^2|\widehat{k}(\omega)|^2\right)$ *for input-gradient of the kernel scales with the dataset size. Yet, we are not interested in the effect of dataset size on the problem and compare models trained on fixed datasets. Hence, For simplicity of exposition, hereafter we assume* $n = 1$.

According to Remark 4, we expect the spectral decay of the kernel gradient to depend primarily on the Fourier transform of the kernel itself, specifically the $|\widehat{k}(\omega)|^2$ term in Eq. (21).

Remark 5. *Writing the function with Representer theorem, after Remark 4, we have* $f(x) = \alpha_1 k(x, x_1)$, *where* x_1 *is a training sample. For clarity without loss of generality, hereafter, we let* $\alpha_1 = 1$ *and denote our single training sample with* x_t.

We denote the spatial dimension by τ, along which we are to use a Fourier basis. Hence, assuming the training sample is a continuous function along the spatial dimension, we denote it by $x_t(\tau)$.

Let x_e denote the sample for which we seek a gradient-based explanation. Let $x_e'(\tau)$ denote the gradient of the kernel w.r.t. the spatial dimension τ:

$$x_e'(\tau) = \nabla_\tau k(x_e(\tau), x_t(\tau)) \tag{22}$$

Remark 6. *We assume the input data exhibits a high degree of spatial input feature correlation both for* $x_t(\tau)$ *and* $x_e(\tau)$. *This can be simply expressed by a high concentration of spatial power spectrum of the input* $|\widehat{x}(\tau)|^2$ *around zero. Therefore, we assume there is a* τ_0 *such that the following condition holds for all* τ *and* τ'*:*

$$|\tau| < \tau_0 \text{ and } |\tau'| > \tau_0 \rightarrow |\widehat{x}(\tau)|^2 \gg |\widehat{x}(\tau')|^2 \tag{23}$$

This assumption is common for image data, where neighboring pixels tend to be highly correlated, see Appendix G for further discussion.

Furthermore, a condition about x_t and x_e, is the existence of an intersection at a certain point in spatial domain. Let $\Delta(\tau) := x_t(\tau) - x_e(\tau)$, then we can express this condition compactly as:

$$\exists \tau^* \quad \text{such that} \quad \Delta(\tau^*) = 0, \tag{24}$$

As we will show in the next lemma, this intersection influences the behavior of the kernel's spectral decay.

Lemma 3. *Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a shift-invariant kernel with spatial Fourier transform $\widehat{k}(\omega)$, and let $\nabla_x k(x, x')$ denote its input-gradient. Suppose there exist trajectories $x_t(\tau)$ and $x_e(\tau)$, with high autocorrelation, as stated in Remark 6. The asymptotic decay rate of the spatial power spectrum of the gradient kernel is primarily determined by the intersection condition stated in Eq. (24).*

Proof. We want to take the Fourier transform (with respect to the spatial variable τ) of the derivative

$$x_e'(\tau) = \frac{d}{dx} k\big(x_e(\tau), x_t(\tau)\big),$$

where, by the shift-invariance of k, we may express

$$\frac{d}{dx} k\big(x_e(\tau), x_t(\tau)\big) = \frac{d}{dx} k\big(x_e(\tau) - x_t(\tau)\big).$$

To make the Fourier analysis tractable, we approximate $\Delta(\tau)$ by a linear function in the spatial domain. Two cases arise:

1. **No Intersection:** If there is no τ for which $\Delta(\tau) = 0$, then a linear approximation yields $\Delta(\tau) = \alpha$ with $\alpha \neq 0$. In this degenerate case, no root is present.
2. **Intersection(s) Exist:** There exists at least one τ^* satisfying
$$\Delta(\tau^*) = 0.$$
In a neighborhood of such a point, by a first-order Taylor expansion,
$$\Delta(\tau) \approx \alpha(\tau - \tau^*),$$
for some $\alpha \neq 0$ and τ close to τ^*. This approximation ensures that the linearized Δ has a root at $\tau = \tau^*$, capturing the sharp transition in the kernel.

In the typical setting with many training samples, it is reasonable to assume that such intersections exits, therefore we focus on case 2. According to the definition of Δ, we have

$$\frac{d}{dx} k\big(x_e(\tau) - x_t(\tau)\big) = \frac{d}{dx} k\big(\Delta(\tau)\big).$$

Under the linearization of Δ, we can use a local change of variables $x = \alpha\tau$ (which, is only valid locally due to the linear approximation), we can write

$$\frac{d}{dx} k\big(-\Delta(\tau)\big) = \frac{1}{\alpha} \frac{d}{d\tau} k\big(\alpha(\tau - \tau^*)\big).$$

Taking the Fourier transform with respect to τ then yields

$$\mathcal{F}_\tau \{x_e'(\tau)\} = \frac{1}{\alpha} \mathcal{F}_\tau \left\{ \frac{d}{d\tau} k\big(\alpha(\tau - \tau^*)\big) \right\}.$$

By the standard property of the Fourier transform, namely that differentiation corresponds to multiplication by $i\omega$, it follows that

$$\mathcal{F}_\tau \left\{ \frac{d}{d\tau} k\big(\alpha(\tau - \tau^*)\big) \right\} = i\omega\, \mathcal{F}_\tau \left\{ k\big(\alpha(\tau - \tau^*)\big) \right\}.$$

Thus, taking the magnitude squared, to compute the power spectrum, we obtain

$$\left| \mathcal{F}_\tau \left\{ \frac{d}{d\tau} k\big(\alpha(\tau - \tau^*)\big) \right\} \right|^2 = \omega^2 \left| \mathcal{F}_\tau \left\{ k\big(\alpha(\tau - \tau^*)\big) \right\} \right|^2.$$

Hence,

$$\left| \mathcal{F}_\tau \{x_e'(\tau)\} \right|^2 = \frac{1}{\alpha^2} \omega^2 \left| \mathcal{F}_\tau \left\{ k\big(\alpha(\tau - \tau^*)\big) \right\} \right|^2.$$

Due to the translation invariance of the Fourier transform, the shift by τ^* does not alter the decay properties, so that

$$\left| \mathcal{F}_\tau \left\{ k\big(\alpha(\tau - \tau^*)\big) \right\} \right|^2 = \left| \mathcal{F}_\tau \left\{ k\big(\alpha\tau\big) \right\} \right|^2.$$

Recalling that the spatial Fourier transform of k is $\hat{k}(\omega)$, we deduce

$$\left| \mathcal{F}_\tau \{x_e'(\tau)\} \right|^2 = \frac{\omega^2}{\alpha^2} \left| \hat{k}(\omega) \right|^2,$$

or, equivalently,

$$\left| \mathcal{F}_\tau \{x_e'(\tau)\} \right|^2 \in \mathcal{O}\Big(\omega^2\, \hat{k}(\omega)^2\Big).$$

Finally, while one might consider the possibility of multiple intersections (i.e., multiple neighborhoods where $\Delta(\tau)$ changes sign), these contribute only as a multiplicative factor in the intermediate expressions (analogous to summing over intersections) and do not affect the order of decay rate. Therefore, the presence of at least one intersection governs the tail behavior of the spatial power spectrum of the gradient kernel. □

Thus, we have demonstrated that the gradient of the samples can be approximated using local linear projections of the gradient of the kernel into the spatial domain. Consequently, the sharp transitions in the spatial domain are a direct consequence of sharp transitions in the gradient of the kernel. Since the tail behavior of the power spectrum depends only on the existence of such sharp transitions and not on their number, we considered a single intersection for a single sample in our analysis.

We now present the proof of Theorem 1 under the assumption that the training and explanation trajectories intersect. This assumption is made primarily for theoretical convenience. In practice, factors beyond our theoretical model—such as random initialization—can induce sharp transitions at arbitrary locations. Consequently, the absence of such intersections is highly improbable when the explained sample lies within the support of the training data distribution. This also aligns with previous findings that sharp transitions induced by ReLU breakpoints, which introduce nonlinearity, occur not only on the training data [43], but also in surrounding regions.

Theorem 1. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a fixed dataset and let $f : \mathbb{R}^d \to \mathbb{R}$ be a neural network whose associated Neural Tangent Kernel (NTK) is denoted by $K^{(NTK)}(c)$. Then, the asymptotic decay of the power spectrum of $K^{(NTK)}(c)$ is directly proportional to the asymptotic decay of the power spectrum of the spatial Fourier transform of $\nabla f(x)$.*

Proof. By Lemma 2, the spectral decay of the input-gradient $\nabla f(x)$ of a shift-invariant kernel is characterized by

$$|\mathcal{F}\{\nabla f\}|^2 = \mathcal{O}(n\,\omega^2 |\hat{k}(\omega)|^2),$$

when focusing on a single training sample and a corresponding explanation instance we realize that n would be a constant.

Next, under the high-autocorrelation assumption and the existence of an intersection between the training and explanation trajectories (as specified in Eq. (24)), Lemma 3 shows that the tail behavior of the spatial power spectrum of the gradient kernel is governed by the local behavior at this intersection. In this region, a linear approximation of the difference $\Delta(\tau)$ is valid, and the induced local (approximate) change of variables

$$x = \alpha\,\tau$$

(with $\alpha \neq 0$) enables us to relate the derivative with respect to x to that with respect to τ via $\frac{d}{dx} = \frac{1}{\alpha}\frac{d}{d\tau}$.

Consequently, the Fourier transform of the gradient undergoes the transformation

$$\mathcal{F}_\tau\left\{\frac{d}{dx}k(\cdot)\right\} = \frac{1}{\alpha}\, i\omega\, \mathcal{F}_\tau\left\{k(\alpha(\tau - c))\right\} \tag{25}$$

$$= \frac{\omega^2}{\alpha^2}\left|\hat{k}(\omega)\right|^2 \tag{26}$$

which implies that the power spectrum is asymptotically proportional to

$$\mathcal{F}_\tau\left\{\frac{d}{dx}k(\cdot)\right\} \in \mathcal{O}\Big(\omega^2\,|\hat{k}(\omega)|^2\Big).$$

Thus, the asymptotic decay of the power spectrum of $K^{(\mathrm{NTK})}(c)$ is directly proportional to that of the spatial Fourier transform of $\nabla f(x)$, as claimed. □

Derivations for Eq. (7) In this section, we compute the two integrals related to the Neural Tangent Kernel (NTK) connections, under the assumption that the underlying kernel is the Laplace kernel [20]. Since the domain under consideration is finite, we evaluate the integrals over a bounded interval (l, h).

Our goal is to establish the relation $\mathcal{G}(f, \widehat{f}) \sim \Delta\,\mathrm{EF}$, by analyzing the asymptotic behavior of both quantities with respect to the kernel variance parameter b, as it gets modified according to Lemma 1.

From Theorem 1, we know that the quantity S_{e_f} in Eq. (1) is asymptotically equivalent to the power spectral density of the Laplace kernel:

$$|\widehat{k}(\omega)|^2 = \frac{2b}{1 + b^2\omega^2} \tag{27}$$

Moreover, Lemma 2 shows that $\mathcal{G}(f, \widehat{f})$ in Eq. (7) has the same leading-order behavior.

Substituting into the integrals and analyzing their asymptotics yields

$$\mathcal{G}(f, \widehat{f}) \sim \Delta\,\mathrm{EF} \sim \mathcal{O}\left(\frac{1}{b}\right), \tag{28}$$

as $b \to \infty$, and

$$\mathcal{G}(f, \widehat{f}) \sim \Delta\,\mathrm{EF} \sim \mathcal{O}(b), \tag{29}$$

as $b \to 0^+$, thus confirming the scaling relation of interest.

We note that alternative derivations are possible, but we chose the most direct approach, which also ensures that the two orthogonal components $(\mathrm{EF} + \Delta\,\mathrm{EF})$ share consistent units.

F. Contribution of ReLU to NTK's Sharpness

The literature of NTK is somewhat denser in terms of the results around ReLU networks, and as far as the authors are concerned, SoftPlus has not been considered as an option when analyzing properties of NTK. Here, we introduce a creative technique to bypass tedious steps for finding analytical solutions to for NTK, assuming that we have some initial results, which is based of convolution operation defined as.

$$f * g(u) = \int_{v\in\mathbb{R}} f(u - v)g(v)dv \tag{30}$$

We focus on a work that provides a simple introduction into the equations needed to be solved for the ReLU NTK.

Summarizing one of the results in [52] for 1 hidden-layer neural networks with initialization weights and bias variances $\sigma_w^2 = 1$, $\sigma_b^2 = 0$, we have to compute the τ-transform, defined as:

$$\tau_\phi\,(c; p) = \mathbb{E}_{z_1, z_2 \sim p(z_1, z_2; c, \sigma^2)}\left[\phi\,(z_1)\,\phi\,(z_2)\right] \tag{31}$$

where ϕ denotes the activation function, and

$$p(z_1, z_2; c, \sigma^2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & c\sigma^2 \\ c\sigma^2 & \sigma^2 \end{bmatrix}\right). \tag{32}$$

After computing the τ-transform, we can compute the NTK using the following recursive equations:

$$K^{(\mathrm{NTK})}(c) = K^{(0)}(c) + cK^{(1)}(c) \tag{33}$$

where $K^{(0)}(c)$ and $K^{(1)}(c)$ are defined as follows

$$K^{(0)}(c) = \tau_{\phi}(c;p) \tag{34}$$
$$K^{(1)}(c) = \tau_{\phi'}(c;p) \tag{35}$$

where $\tau_{\phi'}$ is defined as follows

$$\tau_{\phi'}(c;p) = \frac{\partial_c}{\sigma^2}\tau_{\phi}(c;p)$$

Lemma 1. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be an activation function, and let $K^{(NTK)}(c)$ denote the Neural Tangent Kernel (NTK) associated with ϕ. Define $\phi_\beta = \phi * g_\beta$ as the activation function obtained by convolving ϕ with a Gaussian function $g_\beta(x) = \sqrt{\frac{\beta}{2\pi}}e^{-x^2\beta/2}$ of precision β. Then, the NTK corresponding to ϕ_β, denoted as $K_\beta^{(NTK)}(c)$, leads to a smoother function in the sense that it exhibits faster decay, compared to $K^{(NTK)}(c)$.*

Proof. We can simply start with the definition of τ-transform and the convolution of a Gaussian function g with ReLU ϕ, as follows

$$\tau_{\phi_\beta}(c;p) = \tau_{\phi * g_\beta}(c;p) \tag{36}$$
$$= \iint \phi_\beta(z_1)\phi_\beta(z_2)p(z_1,z_2)dz_1dz_2 \tag{37}$$
$$= \iint \phi * g_\beta(z_2)\phi * g_\beta(z_1)$$
$$p(z_1,z_2)dz_1dz_2 \tag{38}$$
$$= \iiiint \phi(z_2-\nu_2)\phi(z_1-\nu_1)$$
$$g_\beta(\nu_1)g_\beta(\nu_2)p(z_1,z_2)d\nu_1d\nu_2dz_1dz_2 \tag{39}$$
$$= \iiiint \phi(\kappa_1)\phi(\kappa_2)g_\beta(\nu_2)g_\beta(\nu_1)$$
$$p(\kappa_1+\nu_1,\kappa_2+\nu_2)d\kappa_1d\kappa_2d\nu_1d\nu_2 \tag{40}$$
$$= \iint \phi(\kappa_1)\phi(\kappa_2)$$
$$\left(\iint g_\beta(\nu_2)g_\beta(\nu_1)p(\kappa_1+\nu_1,\kappa_2+\nu_2)d\nu_1d\nu_2\right)$$
$$d\kappa_1d\kappa_2 \tag{41}$$
$$= \iint \phi(\kappa_1)\phi(\kappa_2)q(\kappa_1,\kappa_2)d\kappa_1d\kappa_2 \tag{42}$$
$$= \tau_\phi(c;q) \tag{43}$$

This shows that the convolution of ReLU with a Gaussian function changes the covariance matrix of the τ-transform to:

$$q(\boldsymbol{\kappa}) = \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2+\beta & c\sigma^2\\ c\sigma^2 & \sigma^2+\beta\end{bmatrix}\right). \tag{44}$$

As the range of $c \in [-1,1]$ is fixed, increasing β would lead to a matrix closer to identity, hence a smoother kernel.

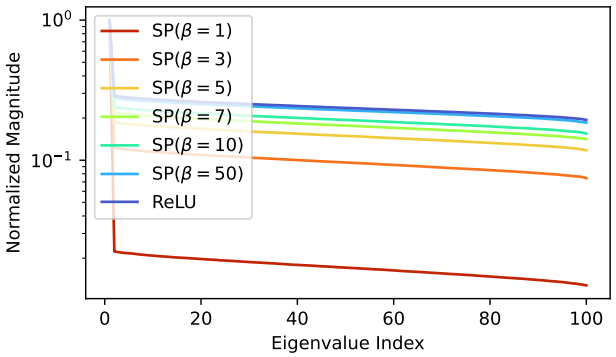


Figure 11. **Spectral Decay of the Empirical NTK Across Smooth Parameterizations of ReLU.** This figure depicts the spectral decay of the empirical neural tangent kernel NTK, plotting the eigenvalues (x-axis) against their normalized magnitudes (y-axis). This empirical evaluation supports Lemma 1, demonstrating how the choice of activation function influences the spectral tail behavior. Notably, ReLU exhibits the heaviest tail, while increasing the Softplus β parameter, shown by SP(β), results in a sharper decay. For further insights, compare these results with Fig. 2 where there is a very similar progression in the tail of the spatial power spectrum. This figure is produced by the code available in [42].

We would like to conclude the proof by highlighting the fact that

$$\frac{d}{dz}(\phi * g_\beta(z)) = \phi' * g_\beta(z) \tag{45}$$

where $\phi' = \frac{d}{dz}\phi$. Therefore, the τ-transform applied to the second term of the kernel, *i.e.* $K^{(1)}$, would lead to the same derivations, with a consistent replacement of $\phi \to \phi'$. □

It is important to note that in practice, we approximate the convolution of Gaussian and ReLU with SoftPlus activation function. That is

$$\phi_\beta(x,\beta) = \phi * g_\beta(x) \approx \text{SoftPlus}(x;\beta), \tag{46}$$

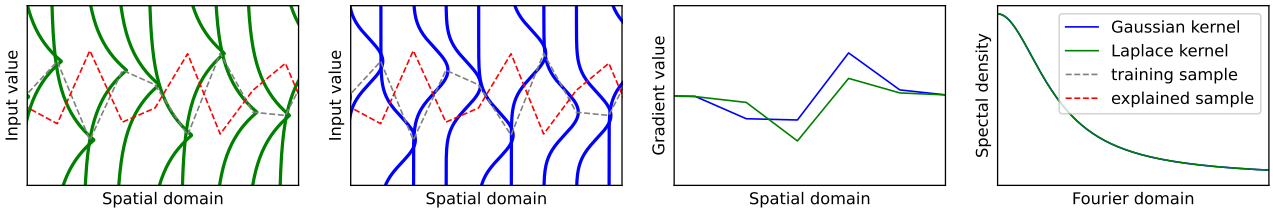
with a proper choice of precision β, which is directly proportional to the parameter of SoftPlus.

We have also verified the statement in Lemma 1 experimentally in Fig. 11, using the method in [42] for SoftPlus activation function.

G. Regarding Input's Spatial Autocorrelation

There are important confounding factors in our analysis, regarding the autocorrelation of the learned features and that of the input features of the data. If the input data has high spatial autocorrelation, then we naturally expect a high autocorrelation in the learned features after training.

Even though being intuitive, specially in image data, this phenomenon has not been theoretically proven yet.



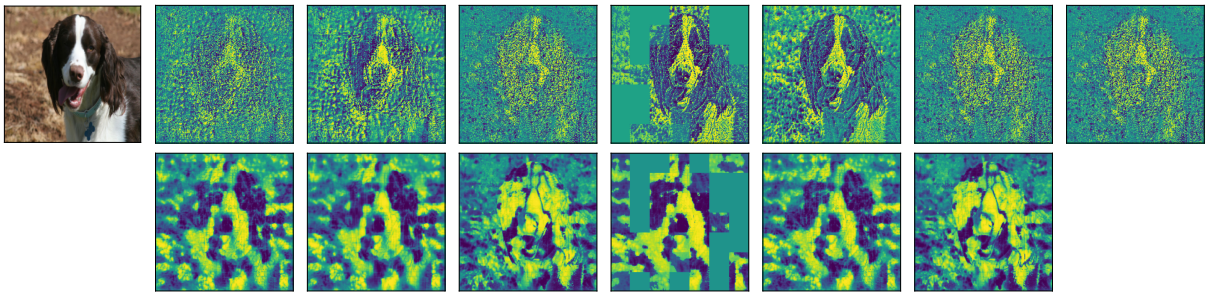
(a) (b) (c) (d)

Figure 12. **Impact of Low Spatial Autocorrelation on Spatial Power Spectrum of Gradient.** This figure illustrates the impact of low spatial autocorrelation in the input and, consequently, in the learned features on the tail behavior of the spatial spectral density of the gradient. Compare with Fig. 10. **(a)** and **(b)**: As in Fig. 10, the x-axis represents spatial dimensions, while the y-axis represents pixel values. However, unlike Fig. 10, where the input and learned features exhibit high spatial autocorrelation, here both display low spatial autocorrelation. This phenomenon can be observed in various data modalities, such as data frames. As a result, the relationship between input samples—specifically, the sample to be explained (red line) and the training sample (gray line)—appears more irregular. **(c)**: Taking the gradient of the classifier with respect to the input yields a function in the spatial domain. Compared to Fig. 10, the gradient function here is less structured, reflecting the reduced spatial autocorrelation. **(d)**: In this case, applying the Fourier transform to the gradient values may not reveal distinct spectral decay rates for the Laplace and Gaussian kernels, as the tail behavior is also influenced by the input autocorrelation.

Unfortunately, proving this property is outside the scope of explainability and is more towards the literature around kernel methods or feature alignment. This is important as we assume continuity and high spatial autocorrelation in our analysis for recovering the behavior of the tail of the power spectrum, without a formal proof.

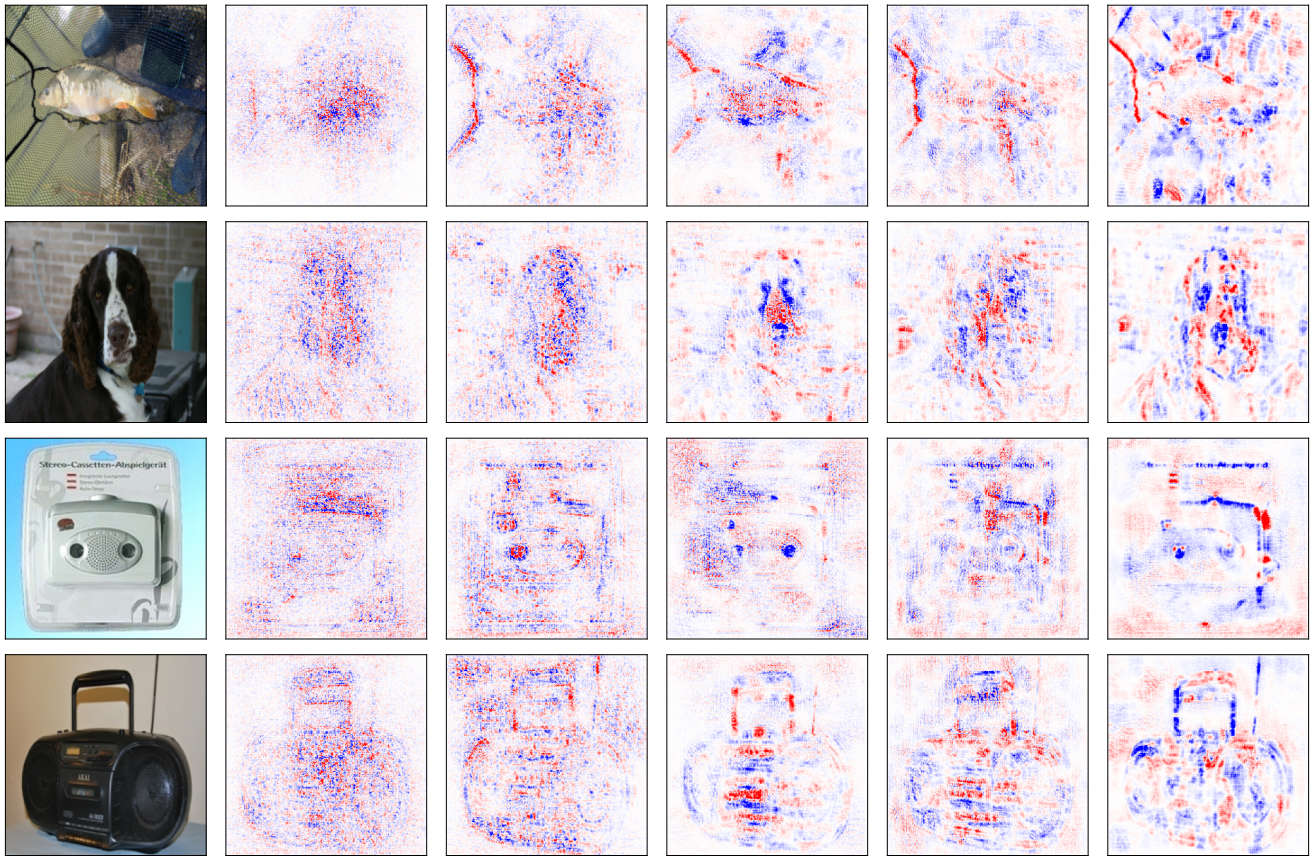
We are aware that such assumption would rule out the existence and influence of sharp changes in the image, such as edges.

For a visualization of low feature autocorrelation that might correspond to a network after initialization, or a network trained on a data with low spatial autocorrelation, compare Figs. 10 and 12.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 13. **Sample Visualization of Explanations Outputs After Inverse Transformation Normalization.** This figure presents a visualization of the outputs from various explanation methods after applying the inverse transformation method. This normalization technique aligns different distributions while preserving their spectral properties and maintaining insensitivity to magnitude.



(a) Input (b) ReLU (c) SP($\beta = 7$) (d) SP($\beta = 5$) (e) SP($\beta = 3$) (f) SP($\beta = .9$)

Figure 14. **Sample Visualization of VanillaGrad Explanations Across Smooth ReLU Parameterizations.** This figure presents additional samples from trained networks with different smooth parameterizations of ReLU, denoted by SP(β). The goal is to illustrate how the reliance on high-frequency information in a ReLU network and its smooth variants manifests in VanillaGrad explanations. This phenomenon has been discussed in more detail in Sec. 3.4. VanillaGrad explanations provide a local snapshot of the network's dependence on features across different frequencies, also known as harmonics [29, 40]. Some of these harmonics can be "suppressed" through surrogates introduced by explanation methods. To ensure an unaltered view of this effect, all visualizations in this figure utilize the simplest gradient-based explanation method, VanillaGrad, applied *without any surrogates*.