Explaining Image Classifiers with Multiscale Directional Image Representation

Stefan Kolek[1], Robert Windesheim[1], Hector Andrade-Loarca[1], Gitta Kutyniok[1,2], Ron Levie[3]
[1]Ludwig-Maximilians-Universität München, Department of Mathematics
[2]University of Tromsø, Department of Physics and Technology
[3]Technion-Israel Institute of Technology, Department of Mathematics
{kolek,windesheim,andrade,kutyniok}@math.lmu.de, levieron@technion.ac.il

Abstract

Image classifiers are known to be difficult to interpret and therefore require explanation methods to understand their decisions. We present ShearletX, a novel mask explanation method for image classifiers based on the shearlet transform – a multiscale directional image representation. Current mask explanation methods are regularized by smoothness constraints that protect against undesirable fine-grained explanation artifacts. However, the smoothness of a mask limits its ability to separate fine-detail patterns, that are relevant for the classifier, from nearby nuisance patterns, that do not affect the classifier. ShearletX solves this problem by avoiding smoothness regularization all together, replacing it by shearlet sparsity constraints. The resulting explanations consist of a few edges, textures, and smooth parts of the original image, that are the most relevant for the decision of the classifier. To support our method, we propose a mathematical definition for explanation artifacts and an information theoretic score to evaluate the quality of mask explanations. We demonstrate the superiority of ShearletX over previous mask based explanation methods using these new metrics, and present exemplary situations where separating fine-detail patterns allows explaining phenomena that were not explainable before.

1. Introduction

Modern image classifiers are known to be difficult to explain. Saliency maps comprise a well-established explainability tool that highlights important image regions for the classifier and helps interpret classification decisions. An important saliency approach frames saliency map computation as an optimization problem over masks [8, 10, 13, 14, 18, 24, 29]. The explanation mask is optimized to keep only parts of the image that suffice to retain the classification decision. However, Fong and Vedaldi [14] showed that an unregularized explanation mask is very susceptible to explanation artifacts and is hence unreliable.

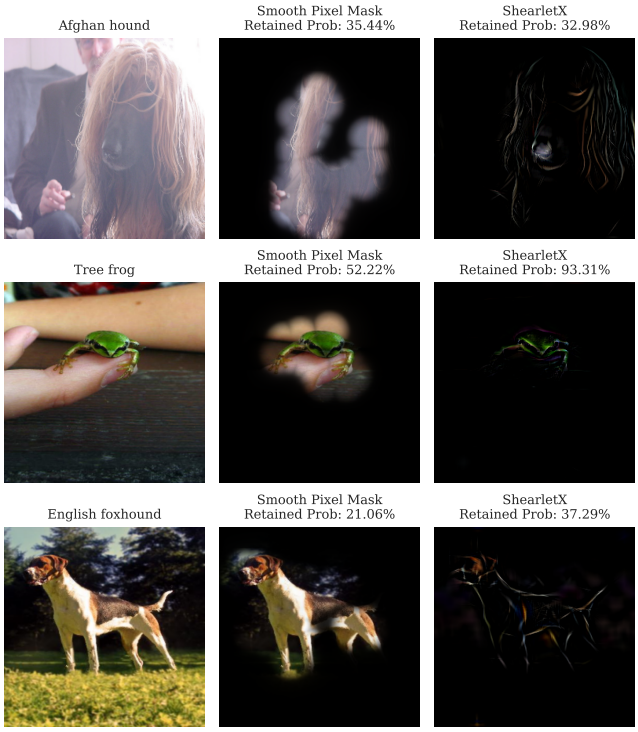

Figure 1. Left column: ImageNet samples with prediction. Middle column: Smooth pixel mask explanation from Fong et al. [13]. Right column: ShearletX (ours). Retained probability is computed as class probability after masking divided by class probability before masking. ShearletX is the first mask explanation method that can separate fine-detail patterns, that are relevant for the classifier, from nearby patterns that are irrelevant, without producing artifacts.

Therefore, current practice [8, 13, 14] heavily regularizes the explanation masks to be smooth. The smooth explanation masks can communicate useful explanatory information by roughly localizing the relevant image region. However, the pattern that is relevant for the classifier is often overlaid on patterns that do not affect the classifier. In such a situation the mask cannot effectively separate the relevant pattern from the nuisance pattern, due to the smoothness con-

straints. As a result, many details that are irrelevant to the classifier, such as background elements, textures, and other spatially localized patterns, appear in the explanation.

An ideal mask explanation method should be resistant to explanation artifacts and capable of highlighting only relevant patterns. We present such a method, called *ShearletX*, that is able to separate different patterns that occupy nearby spatial locations by optimizing a mask in the shearlet representation of an image [25]. Due to the ability of shearlets to efficiently encode directional features in images, we can separate relevant fine-grained image parts, like edges, smooth areas, and textures, extremely well. We show both theoretically and experimentally that defining the mask in the shearlet domain circumvents explanation artifacts. The masked image is optimized so that the classifier retains its prediction as much as possible and to have small spatial support (but not high spatial smoothness), while regularizing the mask to be sparse in the shearlet domain. This regularization assures that ShearletX retains only relevant parts, a fact that we support by a new information theoretic score for the quality of mask explanations. Figure 1 gives examples demonstrating that ShearletX can separate relevant details from nuisance patterns, which smooth pixel masks cannot. Our contributions are summarized as follows:

1. ShearletX: The first mask explanation method that can effectively separate fine-detail patterns, that are relevant for the classifier, from nearby nuisance patterns, that do not affect the classifier.
2. Artifact Analysis: Our explanation method is based on low-level vision for maximal interpretability and belongs to the family of methods that produce out-of-distribution explanations. To validate that the resulting out-of-distribution explanations are meaningful, we develop a theory to analyze and quantify explanation artifacts, and prove that ShearletX is resilient to such artifacts.
3. Hallucination Score: a new metric for mask explanations that quantifies explanation artifacts by measuring the amount of edges in the explanation that do not appear in the original image.
4. Concisesness-Preciseness Score: A new information theoretic metric for mask explanations that gives a high score for explanations that extract the least amount of information from the image to retain the classification decision as accurately as possible.
5. Experimental Results: We demonstrate that ShearletX performs better than previous mask explanations using our new metrics and give examples where ShearletX allows to explain phenomena that were not explainable with previous saliency methods.

The source code for the experiments is publicly available [1].

[1] https://github.com/skmda37/ShearletX

2. Related Work

The explainability field has experienced a surge in research activity over the last decade, due to the societal need to explain machine learning models. We focus on explainability aspects of image classifiers, where saliency maps provide an important and useful way of understanding a classifier's prediction. The community has also introduced other tools, such as concept-based methods [22] and inherently interpretable architectures [9], but we will not focus on these in our work. In the following, we review previously introduced saliency map methods.

Pixel Attribution Methods

Many saliency map methods assign a relevance score to each pixel indicating its relevance for the prediction. Such methods include Gradient Saliency [38], Grad-CAM [37], LRP [5], Guided Backprop [41], and Integrated Gradients [42]. Although these methods can help explain classifiers, they are heuristic in their approach and not optimized for a well-defined notion of relevance. Therefore, the fidelity of pixel attribution methods needs to be checked post-hoc with metrics such as the area over the perturbation curve [4] and can be low. Moreover, Kindermans et al. [1] showed that pixel attribution methods can be highly unreliable. Other well-known explanation methods, such as LIME [35] and SHAP [28] can be applied to images, by first segmenting the image into superpixels and assigning a relevance score to each superpixel. However, research recently revealed various vulnerabilities of LIME and SHAP [40].

Pixel Mask Explanations

Mask explanations do not attribute individual relevance scores to (super)pixels but rather optimize a mask to delete as much information of the image as possible while retaining the classifier's prediction. The advantage of this approach is that one optimizes for a natural interpretability objective that can be quickly validated in two steps: (1) Determining which and how much information was deleted by the mask (2) Computing the class probability score after masking the image. Fong and Vedaldi [14] were the first to find an explanation mask as a solution to an optimization problem that can be summarized as

$$\max_{m \in \mathcal{M}} \; \mathop{\mathbb{E}}_{u \sim \nu} \Big[\Phi_c(x \odot m + (1 - m) \odot u) \Big] - \lambda \cdot \|m\|_1, \quad (1)$$

where $x \in \mathbb{R}^d$ is the input image, Φ_c returns the classifier's class probability, $u \in \mathbb{R}^d$ is a random perturbation from a predefined probability distribution ν (*e.g.*, constant, blur, or noise), $m \in \mathbb{R}^d$ is a mask on x, $\lambda \in \mathbb{R}_+$ is the Lagrange multiplier encouraging sparsity in m, and $\mathcal{M}$ is a prior over the explanation masks. Fong and Vedaldi [14] found that

not choosing a prior, *i.e.* $\mathcal{M} = [0,1]^d$, produces explanation artifacts. To mitigate artifacts, they enforce a more regular structure on the mask by using an upsampled lower resolution mask and regularizing the mask's total variation (TV). Fong et al. [13] improved this method by reformulating the area constraint and adding a new parametric family of smooth masks, which allowed to remove all hyperparameters from the optimization problem. The masks remain extremely smooth but the main advantage is that the size of the mask can be controlled by an area constraint that directly controls the size of the mask as a percentage of the total image area. We will refer to this method as *smooth pixel mask* to highlight the fact that this method produces extremely smooth explanations due to strong smoothness constraints on the mask.

Wavelet Mask Explanations

Kolek et al. [24] proposed the *CartoonX* method, which masks in the wavelet representation of images to extract the relevant piece-wise smooth part of an image. Wavelets sparsely represent piece-wise smooth images and therefore the wavelet sparsity constraint in CartoonX typically leads to piece-wise smooth explanations. However, Kolek et al. [24] do not compare CartoonX to smooth pixel masks [13], which also enforce piece-wise smoothness by regularizing and parameterizing the pixel mask. Besides lacking a clear advantage over smooth pixel masks, we find that CartoonX produces blurry spatial areas that can be quite difficult to interpret (see Figure 2). ShearletX improves upon CartoonX by (a) leveraging the advantages of shearlets over wavelets for representing edges in images, (b) eliminating an ambiguous spatial blur in CartoonX, and (c) having a clear advantage over smooth pixel masks.

3. Background

To develop and analyze ShearletX we need to first give the necessary technical background for wavelets [30] and shearlets [25] in the context of images.

Wavelets for Images

A gray-level image can be mathematically modeled as a square integrable function $f : \mathbb{R}^2 \to \mathbb{R}$. A wavelet $\psi : \mathbb{R}^2 \to \mathbb{R}$ is a spatially localized bump with oscillations, that is used to probe the local frequency, or scale, of an image. Three suitably chosen mother wavelets $\psi^1, \psi^2, \psi^3 \in L^2(\mathbb{R}^2)$ with dyadic dilations and translations yield an orthonormal basis

$$\left\{\psi^k_{j,n} := \frac{1}{2^j}\psi^k\left(\frac{\cdot - 2^j n}{2^j}\right)\right\}_{j\in\mathbb{Z}, n\in\mathbb{Z}^2, 1\le k\le 3} \tag{2}$$

of the square integrable function space $L^2(\mathbb{R}^2)$. The three indices $k \in \{1,2,3\}$ correspond to vertical, horizontal, and diagonal directions. The image f can be probed in direction $k \in \{1,2,3\}$, at location $n \in \mathbb{Z}^n$, and at scale $2^j \in \mathbb{Z}$ by taking the inner product $\langle f, \psi^k_{j,n}\rangle$, which is called a *wavelet (detail) coefficient*. The wavelet coefficient $\langle f, \psi^k_{j,n}\rangle$ has high amplitude if the image f has sharp transitions over the support of $\psi^k_{j,n}$. Pairing $\psi^1, \psi^2, \psi^3 \in L^2(\mathbb{R}^2)$, with an appropriate scaling function $\phi \in L^2(\mathbb{R}^2)$, defines a multiresolution approximation. More precisely, for all $J \in \mathbb{Z}$, any finite energy image f decomposes into

$$f = \sum_{n\in\mathbb{Z}^2} a_n \phi_{J,n} + \sum_{1\le k\le 3}\sum_{j\le J,} d^k_{j,n}\psi^k_{j,n}, \tag{3}$$

where $a_n = \langle f, \phi_{J,n}\rangle$ and $d^k_{j,n} = \langle f, \psi^k_{j,n}\rangle$ are the approximation coefficients at scale J and wavelet coefficients at scale $j-1$, respectively. In practice, images are discrete signals $x[n_1, n_2]$ with pixel values at discrete positions $n = (n_1, n_2) \in \mathbb{Z}^2$ but they can be associated with a function $f \in L^2(\mathbb{R}^2)$ that is approximated at some scale 2^L by x. The discrete wavelet transform (DWT) of an image x then computes an invertible wavelet image representation

$$\mathcal{DWT}(x) = \left\{a_{J,n}\right\}_n \cup \left\{d^1_{j,n}, d^2_{j,n} d^3_{j,n}\right\}_{L<j\le J,n} \tag{4}$$

corresponding to discretely sampled approximation and wavelet coefficients of f.

Shearlets for Images

Wavelets are optimal sparse representations for signals with point singularities [12], in particular, piece-wise smooth 1d signals. However, images are 2d signals where many singularities are edges, which are anisotropic (directional), and are not optimally represented by wavelets. Shearlets [25] extend wavelets and form a multiscale directional representation of images, which allows efficient encoding of anisotropic features. Next, we describe the continuous shearlet system, and note that the discrete shearlet system is just a discrete sampling of the continuous system. The shearlet transform was introduced in [16]. Similarly to the wavelet transform, the shearlet transform applies transformations to a function, called the mother shearlet, to generate a filter bank. The transformations are (a) translation, to change the location of the shearlet probe, (b) anisotropic dilation, to change the scale and shape, creating elongated probes of different scales, and (c) shearing, to probe at different orientations. To dilate and shearing a function, we define the following three matrices:

$$A_a := \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix}, \quad \widetilde{A}_a := \begin{pmatrix} \sqrt{a} & 0 \\ 0 & a \end{pmatrix}, \quad S_s := \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix},$$

where $s, a \in \mathbb{R}$. Given $(a, s, t) \in \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}^2$, $\psi \in L^2(\mathbb{R}^2)$, and $x \in \mathbb{R}^2$, we define

$$\begin{aligned} \psi_{a,s,t,1}(x) &:= a^{-\frac{3}{4}} \psi\left(A_a^{-1} S_s^{-1}(x-t)\right), \\ \psi_{a,s,t,-1}(x) &:= a^{-\frac{3}{4}} \widetilde{\psi}\left(\widetilde{A}_a^{-1}(S_s^T)^{-1}(x-t)\right), \end{aligned} \tag{5}$$

where $\widetilde{\psi}(x_1, x_2) := \psi(x_2, x_1)$, for all $x = (x_1, x_2) \in \mathbb{R}^2$, and ψ is the mother shearlet. The continuous shearlet transform is then defined as follows.

Definition 3.1 (Continuous Shearlet Transform)**.** Let $\psi \in L^2(\mathbb{R}^2)$. Then the family of functions $\psi_{a,s,t,\iota} \colon \mathbb{R}^2 \to \mathbb{R}$ parametrized by $(a, s, t, \iota) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^2 \times \{-1, 1\}$ that are defined in (5) is called a *shearlet system*. The corresponding shearlet transform is defined by

$$\mathcal{SH}_\psi : L^2(\mathbb{R}^2) \to L^\infty\big(\mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^2 \times \{-1, 1\}\big), \tag{6}$$

where $\mathcal{SH}_\psi(f)(a, s, t, \iota) := \langle f, \psi_{a,s,t,\iota} \rangle$.

The continuous shearlet transform can be digitized to the *digital shearlet transform*[2] [27], denoted as $\mathcal{DSH}$, by sampling a discrete system from the function system (6). Note that the digital shearlet transform, like the discrete wavelet transform, is an invertible transformation.

4. Method

In this section, we develop our novel mask explanation method *ShearletX* (Shearlet Explainer).

ShearletX

The optimization objective for ShearletX is

$$\begin{aligned} \max_m \; & \mathop{\mathbb{E}}_{u \sim \nu} \Big[\Phi_c(\mathcal{DSH}^{-1}(m \odot \mathcal{DSH}(x) + (1-m) \odot u)) \Big] \\ & - \lambda_1 \|m\|_1 - \lambda_2 \|\mathcal{DSH}^{-1}(m \odot \mathcal{DSH}(x))\|_1, \end{aligned} \tag{7}$$

where $m \in [0, 1]^n$ denotes a mask on the digital shearlet coefficients, Φ_c returns the class probability of the classifier, ν is the perturbation distribution, $\lambda_1 \in \mathbb{R}_+$ controls the sparseness of the shearlet mask, and $\lambda_2 \in \mathbb{R}_+$ controls the penalty for spatial energy. The final ShearletX explanation is given by masking the shearlet coefficients and inverting the masked shearlet coefficients back to pixel space, *i.e.*,

$$\text{ShearletX}(x) := \mathcal{DSH}^{-1}(m \odot \mathcal{DSH}(x)). \tag{8}$$

The expectation term in the ShearletX objective (7) ensures that the image after masking and perturbing retains the classification decision. We find that the spatial penalty is a crucial technical addition, that deletes classifier irrelevant spatial energy in the explanation and ensures that no irrelevant blurry areas remain in the explanation, as opposed to CartoonX [24] (see Figure 2). A smooth area is retained by ShearletX only if it is important for the classifier (see English Foxhound and Frog in Figure 1). Moreover, the color can be distorted if the original color is not important. When the color is important for the classifier, ShearletX will keep the color (see, for example, Figure 1, where ShearletX keeps the brown color of the English Foxhound's head and the green color of the Frog).

For the perturbation distribution ν, we deliberately avoid in-distribution perturbations from an in-painting network, as opposed to [8]. The reason is that in-distribution masks may delete parts of the image that are relevant to the classifier if the in-painter in-fills such parts effectively, making the explanation hard to interpret. Therefore, we follow the out-of-distribution approach of CartoonX [24], and use white noise in the representation system that is adapted to the mean and variance of the shearlet coefficients (see Supplementary Material B.1 for details).

[2]For the digital shearlet transform, we used pyshearlab from http://shearlab.math.lmu.de/software#pyshearlab.

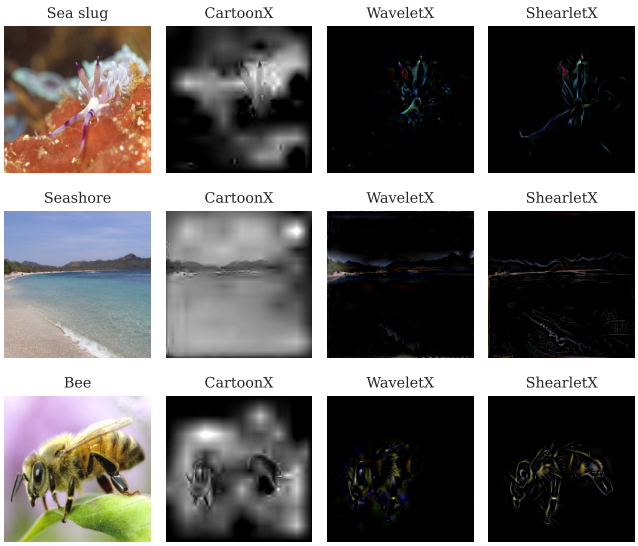

Figure 2. Left Column: Input images classified by VGG-19 [39]. Comparing CartoonX by Kolek et al. [24], WaveletX (ours), and ShearletX (ours). WaveletX improves CartoonX significantly due to the spatial penalty that eliminates undesirable blurry spatial areas that are difficult to interpret. Note that ShearletX represents relevant edges better and produces much crisper explanations than WaveletX. This is because anisotropic features, such as edges, can be encoded more efficiently with shearlets than with wavelets.

WaveletX

Solely adding the spatial penalty to the CartoonX objective yields significantly better explanations than the original CartoonX method and eliminates the undesirable blurry areas, that are difficult to interpret (see Figure 2). We will

refer to this new method as WaveletX to highlight the fact that WaveletX and ShearletX only differ in the choice of the representation system. The WaveletX optimization objective is

$$\max_{m} \mathop{\mathbb{E}}_{u\sim\nu}\left[\Phi_c(\mathcal{DWT}^{-1}(m\odot\mathcal{DWT}(x)+(1-m)\odot u))\right] - \lambda_1\|m\|_1 - \lambda_2\|\mathcal{DWT}^{-1}(m\odot\mathcal{DWT}(x))\|_1, \quad (9)$$

where $\mathcal{DWT}$ denotes the discrete wavelet transform of images. Note that we recover the CartoonX [24] objective if we set $\lambda_2 = 0$. In Figure 2, we compare CartoonX [24], WaveletX, and ShearletX on examples classified by a VGG-19 [39] network trained on ImageNet [11].

5. Theory

Fong and Vedaldi [14] first observed the problem of explanation artifacts for mask explanations. We identify explanation artifacts as artificial edges in the explanation (see Figure 3). Artificial edges can form patterns that activate the class label but are not present in the original image. Therefore, a good mask explanation method should not be able to form many artificial edges. In this section, we show theoretically that ShearletX and WaveletX are not prone to artificial edges, by proving that the continuous counterparts of WaveletX and ShearletX cannot create edges that are not present in the original image.

ShearletX is Resistant to Edge Artifacts

In this section, we prove that ShearletX applied to continuous images cannot create artificial edges. When working with shearlets, it is common to model edges as the *wavefront* set of a continuous image [2,3]. The wavefront set is a concept that characterizes the oriented singularities of distributions, in particular, of $L^2(\mathbb{R}^2)$ functions. We state the mathematical definition of the wavefront set below and provide an intuitive explanation afterwards since the definition is somewhat technical.

Definition 5.1. [20, Section 8.1] Let $f \in L^2(\mathbb{R}^2)$ and $k \in \mathbb{N}$. A point $(x,\lambda) \in \mathbb{R}^2 \times \mathbb{S}^1$ is a *k-regular directed point of f* if there exist open neighbourhoods U_x and V_λ of x and λ, respectively, and a smooth function $\phi \in C^\infty(\mathbb{R}^2)$ with $\operatorname{supp}\phi \subset U_x$ and $\phi(x) = 1$ such that

$$\left|\widehat{\phi f}(\xi)\right| \leq C_k\left(1+|\xi|\right)^{-k} \quad \forall\, \xi \in \mathbb{R}^2 \setminus \{0\} \text{ s.t. } \xi/|\xi| \in V_\lambda$$

holds for some $C_k > 0$, where $\widehat{f}$ denotes the Fourier transform of f. The *k-wavefront set* $\mathrm{WF}_k(f)$ is the complement of the set of all k-regular directed points and the *wavefront set* is defined as $\mathrm{WF}(f) := \bigcup_{k\in\mathbb{N}} \mathrm{WF}_k(f)$.

The wavefront set defines the directional singularities of a function f via the Fourier decay of local patches of the function. For piece-wise smooth images with discontinuities along smooth curves, the wavefront set is exactly the set of edges with the orientation of the edge. This explains why the wavefront set is a good model for edges. The wavefront set of an image can be completely determined by the decay properties of its shearlet coefficients [15]. More precisely, the regular point-direction pairs of an image (the complement of the wavefront set) are exactly the pairs of locations and directions where the shearlet coefficients exhibit rapid decay as $a \to 0$ (the precise statement can be found in Supplementary Material A). We use this property of shearlets to prove that ShearletX cannot produce artificial edges for continuous images.

Theorem 1. *Let $x \in L^2[0,1]^2$ be an image modeled as a L^2-function. Let m be a bounded mask on the shearlet coefficients of x and let $\hat{x}$ be the image x masked in shearlet space with mask m. Then, we have $WF(\hat{x}) \subset WF(x)$ and thus masking in shearlet space does not create new edges.*

The idea behind the proof is that creating artificial singularities in regular point-directions of the image would require creating asymptotically slower shearlet decay by masking the coefficients. This is impossible, as masking can only increase the decay rate. See Supplementary Material A for a full proof of Theorem 1.

While in the real world images are digital, they are still an approximation of continuous images that becomes better with increasing resolution. In Section 7, we show experimentally that Theorem 1 indeed predicts the behavior of masked digital images, and ShearletX is not susceptible to explanation artifacts.

WaveletX is Resistant to Edge Artifacts

When analyzing WaveletX, we opt to model singularities via local Lipschitz regularity instead of using the wavefront set approach. This approach is preferable since the Lipschitz regularity of a function is completely characterized by the rate of decay of its wavelet coefficients, as the scale goes to zero [31, Theorem 9.15]. We hence define a regular point as a point for which the image is α-Lipschitz regular in a neighborhood of the point, with $\alpha \geq 1$ (see Definition A.2 in the supplementary material for Lipschitz regularity, and, in particular, the definition for $\alpha > 1$). A singular point is a point which is not regular. Singular points describe image elements such as edges and point singularities.

Theorem 2 (Informal version of Theorem 6)**.** *Let $x \in L^2[0,1]^2$ be an image modeled as a L^2-function. Masking the wavelet coefficients of x with a bounded mask cannot create new singularities.*

The above theorem is an informal version of our formal Theorem 6 in Supplementary Material A. Similarly to

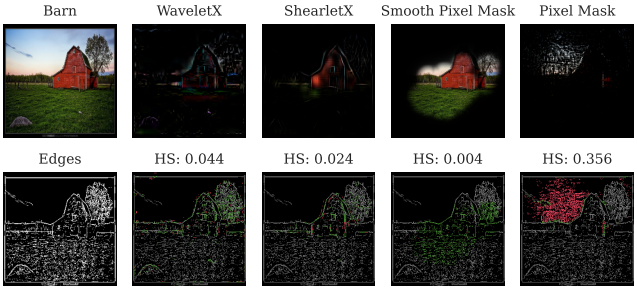

Figure 3. First row: image correctly classified as barn, WaveletX, ShearletX, smooth pixel mask by Fong et al. [14], and pixel mask without smoothness constraints. Second row visualizes the edges in the image and all explanations. Edges marked in red are artificial and quantified by the hallucination score (HS). Green edges are present in the original image. The pixel mask without smoothness constraints hallucinates an artificial barn, which is an example of an explanation artifact and results in a very high HS.

ShearletX, Theorem 2 predicts the behavior for digital images well, and WaveletX is not prone to produce explanation artifacts.

6. Explanation Metrics for Mask Explanations

We now propose two new explanation metrics for mask explanations: (1) The *conciseness-preciseness* (CP) score to evaluate the preciseness of a mask explanation adjusted for its conciseness (2) The *hallucination score* to quantify explanation artifacts.

Conciseness and Preciseness

Several metrics, such as remove and retrain (ROAR) [19] and area over the perturbation curve (AOPC) [4], have been proposed to quantify the fidelity of saliency maps. However, these metrics are designed for pixel attribution methods that provide an ordering of feature importance. Mask explanations can be immediately evaluated by simply plugging in the masked image into the classifier and checking the class probability. A good mask explanation retains the class probability of the prediction. We refer to this property as *preciseness* of the explanation. However, a good explanation mask should not only be precise but also *concise*, *i.e.*, the mask should extract the least amount of information from the available pool of data. We introduce a class of new explanation metrics that combine both aspects into one metric, which we call *conciseness-preciseness* (CP) scores. The definition is

$$\text{CP} = \frac{\text{Retained Class Probability}}{\text{Retained Image Information}}. \tag{10}$$

The retained class probability (preciseness) is computed as the class probability after masking divided by the class probability before masking. We compute the retained information of the image (conciseness) as the information of the masked image divided by the information of the original image. We experiment with three different ways of measuring the information of the image: (1) CP-Entropy: The entropy in the respective image representation system (wavelet, shearlet, or pixel), (2) CP-ℓ_1: The ℓ_1-norm in the respective representation system (wavelet, shearlet, or pixel), (3) CP-ℓ_1 Pixel: The ℓ_1-norm in pixel space irrespective of representation system. For the CP-Entropy, we compute the retained image information of an image with representation coefficients $\{c_i\}_i$ and mask $\{m_i\}_i$ as

$$\exp\left(H\{|m_i c_i|^2\}_i\right) / \exp\left(H\{|c_i|^2\}_i\right) \tag{11}$$

where H denotes the entropy of the induced probability distributions. We use here the exponential of the entropy, also called the *extent* [6], as it balances the "dimensions" and when normalized as in (11) it does not depend on the unit of length in the domain of the image. For CP-ℓ_1, we compute the retained information as the *relative sparsity* $\|\{m_i c_i\}_i\|_1 / \|\{c_i\}_i\|_1$. Note that by measuring information through entropy or ℓ_1-norm in the respective representation system we normalize for the fact that shearlets and wavelets already represent images much more sparsely than pixel representations. The CP score can be interpreted as a measure of preciseness adjusted for by the conciseness of the explanation. Explanations with higher CP scores are superior, assuming no explanation artifacts, which we measure with another metric that we define next.

Hallucination Score

Artificial edges in a mask explanation are edges that are present after masking that were not present in the original image. They can form artificial patterns, that appear as hallucinations (see the far right example in Figure 3). The hallucinations can activate the class label, which is what the explanation optimized for, but do not actually explain the prediction. Therefore, artificial edges are undesirable and can lead to explanation artifacts. We propose to measure such explanation artifacts with a metric that we call *hallucination score* (HS). We compute the hallucination score of an explanation as the number of edges that occur in the explanation but not in the original image, normalized by the number of edges that occur in the original image:

$$\text{HS} = \frac{\#\Big(\text{Edges(Explanation)} \setminus \text{Edges(Image)}\Big)}{\#\text{Edges(Image)}}, \tag{12}$$

where "Explanation" refers to the image obtained after masking, "Image" refers to the original input image, and "Edges" denotes an edge extractor that computes the set of pixels that belong to the edges of the input. Figure 3 provides an example for the hallucination score.

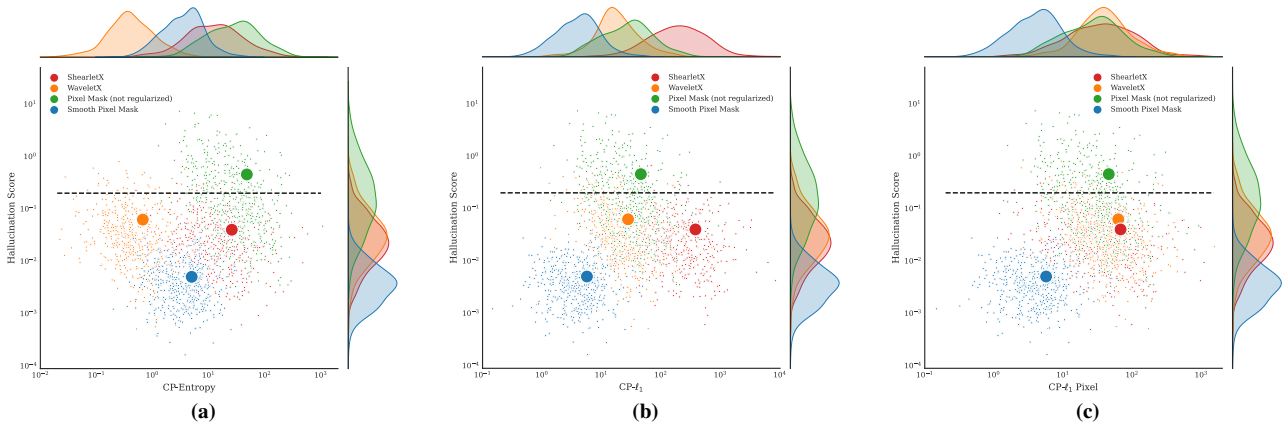

Figure 4. Scatter plot of hallucinaton score (lower is better) and conciseness-preciseness score (higher is better) for ShearletX, WaveletX, smooth pixel masks [13], and pixel mask without smoothness constraints. Retained information of an image for the CP score is measured (a) as entropy in respective representation, (b) as ℓ_1-norm in respective representation, (c) as ℓ_1-norm in pixel space irrespective of representation. The black horizontal line marks explanations where the artificial edges amount to 20% of all edges in the original image. The mean hallucination and CP score are highlighted as big colored dots. ShearletX beats smooth pixel masks and WaveletX across all CP scores while having much better hallucination score than pixel masks without smoothness constraints.

7. Experiments

In this section, we experimentally show (a) that ShearletX and WaveletX do not create a significant amount of artificial edges in practice and are thus resilient to explanation artifacts and (b) that ShearletX outperforms all other mask explanation methods in conciseness-preciseness scores.

Implementation

We use the ImageNet [11] dataset and indicate in each experiment which classification model was used. For ShearletX, we optimize the shearlet mask with the Adam optimizer [23] for 300 steps on the ShearletX objective in (7). We use a single shearlet mask for all three RGB channels as in [24]. The hyperparameter choice for ShearletX, WaveletX, smooth pixel masks, and pixel masks without smoothness constraints are discussed in detail in Supplementary Material B.1. We note as a limitation that, in practice, ShearletX is $5\times$ times slower than smooth pixel masks [13] and WaveletX but not prohibitively slow for many applications (see Supplementary Material B.2 for runtime comparison). For details on the edge detector that we used for the hallucination score, see Supplementary Material B.1.

Comparison of Mask Explanations

We compute ShearletX, WaveletX, the smooth pixel mask by Fong et al. [14] (with area constraint 20%), and the pixel mask without smoothness constraints for 500 random samples from the ImageNet validation dataset and compute the hallucination scores and conciseness-preciseness scores, which are plotted in Figure 4. We use a ResNet18 [17] classifier but our results are consistent across different ImageNet classifiers and different area constraints (5%, 10%, and 20%) for the smooth pixel mask (see Supplementary Material B.3).

The scatter plots in Figure 4 show that pixel masks without smoothness constraints have extremely high hallucination scores, which confirms their proneness to explanation artifacts. The smooth pixel masks by Fong et al. [14] have almost no artificial edges (hallucination score very close to zero) because the masks are constrained to be extremely smooth. ShearletX and WaveletX have on average a moderately higher hallucination score than smooth pixel masks but their upper tail remains vastly lower than the tail for pixel masks without smoothness constraints (note the logarithmic scales in the scatter plots). In Figure 3, one can see that a hallucination score in the order of 10^{-2} produces very few visible artificial edges. Therefore, we conclude from the scatter plot that ShearletX and WaveletX create very few artificial edges and are resilient to explanation artifacts. This also confirms that our Theorem 1 and Theorem 2 approximately hold for discrete images.

Figure 4 further shows that ShearletX has a significantly higher CP-Entropy, CP-ℓ_1, and CP-ℓ_1 Pixel score than the smooth pixel masks by Fong et al. [14]. This validates our claim that ShearletX can delete many irrelevant features that the smooth pixel masks cannot. ShearletX also outperforms WaveletX on all CP scores and even slightly on the hallucination score. Pixel masks without smoothness constraints have the highest CP-Entropy score but are disqualified due to their unacceptable hallucination score. ShearletX, is the only method that has top performance on all CP scores.

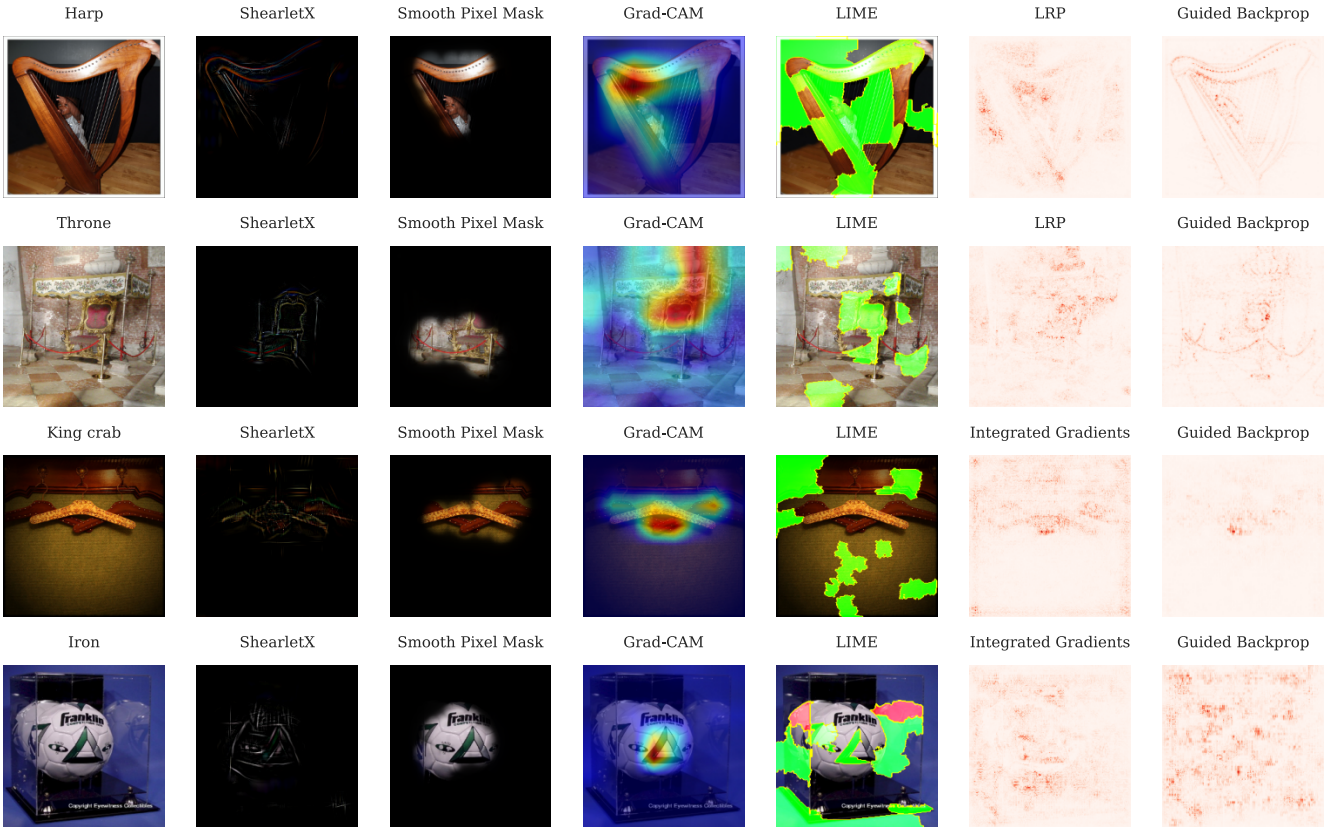

Figure 5. ShearletX compared to smooth pixel mask [13], LIME [35], Grad-CAM [37], LRP [5], Integrated Gradients [42], and Guided Backprop [41]. First two examples are correctly classified by VGG [39] and last two examples are misclassified by MobilenetV3Small [21]. For the harp, note that ShearletX is the only method that effectively seperates the harp from the child playing the harp, indicating, the harp can also be correctly classified without a human playing the harp. For the throne, we observe that ShearletX is able to seperate the throne from other decorations. Here, other methods such as guided backprop, seem to operate more like an edge detector and highlight many other edges such as the floor tiles. Smooth methods such as Grad-CAM and the smooth pixel mask give very rough localizations of part of the throne. Finally, for the misclassifications, which we find more challenging to explain, we can see that only ShearletX can effectively expose the crab and the iron that the classifier saw in the hangers and soccer ball, respectively.

General Saliency Map Comparison

Our experimental results proved that ShearletX has an advantage over the state of the art mask explanation by Fong et al. [13]. Other saliency map methods [5,37,41,42] assign a relevance score to each pixel, allowing to order pixels by relevance. Such methods need to be quantitatively validated post-hoc with metrics such as the area over the perturbation curve [4] or the pointing game [13]. It is challenging to meaningfully compare ShearletX on such metrics, since (1) we cannot order the features in ShearletX by relevance due to the binary nature of masks and (2) in ShearletX the mask is in shearlet space and not in pixel space. Nevertheless, in Supplementary Material B.4 we add a quantitative comparison. In Figure 5, we compare ShearletX qualitatively to pixel attribution methods to demonstrate the insights ShearletX can give that more heuristic pixel attribution methods cannot.

8. Conclusion

We presented ShearletX, a novel mask explanation method, and two explanation metrics (hallucination score and conciseness-preciseness score) to evaluate mask explanations. ShearletX is more effective than other methods at separating fine-detail patterns, which are relevant for the classifier, from nearby nuisance patterns, that do not affect the classifier. Our theoretical results and experiments show ShearletX is well-protected against explanation artifacts and delivers superior explanations than previous mask explanation methods. Our examples illustrate cases when ShearletX can meaningfully interpret classifications but pixel attribution methods cannot. In the future, we will focus on improving the runtime of ShearletX, which is currently slower than other mask explanation methods, to provide real-time explanations of excellent quality.

References

[1] Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 11700:267, 2019. 2

[2] Héctor Andrade-Loarca, Gitta Kutyniok, and Ozan Öktem. Shearlets as feature extractor for semantic edge detection: the model-based and data-driven realm. *Proceedings of the Royal Society A*, 476(2243):20190841, 2020. 5

[3] Héctor Andrade-Loarca, Gitta Kutyniok, Ozan Öktem, and Philipp Christian Petersen. Extraction of digital wavefront sets using applied harmonic analysis and deep neural networks. *SIAM J. Imaging Sci.*, 12:1936–1966, 2019. 5

[4] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. *arXiv:1706.07206*, 2017. 2, 6, 8, 15

[5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 2015. 2, 8

[6] L. L. Campbell. Exponential entropy as a measure of extent of a distribution. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 5(3):217–225, 1966. 6

[7] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 14

[8] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *Proceedings of the 7th International Conference on Learning Representations, ICLR*, 2019. 1, 4

[9] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 2

[10] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 7, 15

[12] Ronald A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998. 3

[13] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 3, 7, 8, 14, 15, 17, 18

[14] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, 2017. 1, 2, 5, 6, 7

[15] Philipp Grohs. Continuous shearlet frames and resolution of the wavefront set. *Monatsh. Math.*, 164(4):393–426, 2011. 5

[16] Kanghui Guo, Gitta Kutyniok, and Demetrio Labate. Sparse multidimensional representations using anisotropic dilation and shear operators. In *Wavelets and Splines*, pages 189–201, Nashville, TN, 2005. Nashboro Press,. 3

[17] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 7, 15, 16, 17

[18] Cosmas Heiß, Ron Levie, Cinjon Resnick, Gitta Kutyniok, and Joan Bruna. In-distribution interpretability for challenging modalities. *Preprint arXiv:2007.00758*, 2020. 1

[19] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9737–9748. Curran Associates, Inc., 2019. 6

[20] Lars Hormander. *The Analysis of Linear Partial Differential Operators. I, Distribution Theory and Fourier Analysis*. Grundlehren Der Mathematischen Wissenschaften. Springer, 1990. 5, 11

[21] Andrew G. Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019. 8, 15, 16, 18

[22] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, 2018. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 7, 14

[24] Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. Cartoon explanations of image classifiers. In *European Conference of Computer Vision (ECCV)*, 2022. 1, 3, 4, 5, 7, 14, 15

[25] G. Kutyniok and D. Labate. *Shearlets: Multiscale Analysis for Multivariate Data*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 2012. 2, 3

[26] G. Kutyniok and D. Labate. *Shearlets: Multiscale Analysis for Multivariate Data*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 2012. 11

[27] Gitta Kutyniok, Wang-Q Lim, and Rafael Reisenhofer. Shearlab 3d: Faithful digital shearlet transforms based on compactly supported shearlets. *ACM Trans. Math. Softw.*, 42(1), jan 2016. 4

[28] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. 2

[29] Jan Macdonald, Stephan Wäldchen, Sascha Hauch, and Gitta Kutyniok. A rate-distortion framework for explaining neural network decisions. *Preprint arXiv:1905.11092*, 2019. 1

[30] Stéphane Mallat. *A wavelet tour of signal processing (2. ed.)*. Academic Press, 1999. 3

[31] Stphane Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, Inc., USA, 3rd edition, 2008. 5, 11

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *arXiv:1912.01703*, 2019. 13

[33] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018. 15

[34] Rafael Reisenhofer and Emily J. King. Edge, ridge, and blob detection with symmetric molecules. *SIAM Journal on Imaging Sciences*, 12(4):1585–1626, 2019. 14

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. 2, 8, 15

[36] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28:2660–2673, 2017. 15

[37] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2019. 2, 8, 15, 16

[38] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Preprint arXiv:1312.6034*, 2014. 2

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 4, 5, 8, 15, 16

[40] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2020. 2

[41] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. 2, 8, 15, 16

[42] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, volume 70, page 3319–3328, 2017. 2, 8, 15, 16