Explaining Image Classifiers with Multiscale Directional Filters

Anonymous CVPR submission

Paper ID *****

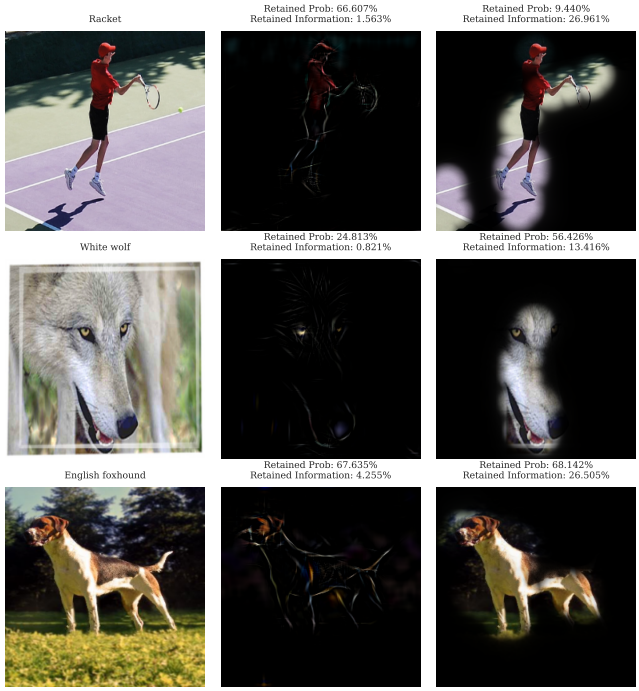


Figure 1. Left column: Imagenet images with VGG-19 [34] prediction. Middle column: Our new ShearletX explanation. Right column: Smooth pixel mask explanation from Fong et al. [11]. Our mask explanation method can delete irrelevant features such as background elements and texture when they are deemed irrelevant without producing explanation artifacts.

Abstract

We present a novel mask explanation method ShearletX that can localize and delete irrelevant features that current mask explanation methods cannot due to their inflexible smoothness constraints that protect against undesirable artifacts. Our method optimizes a sparse mask in the shearlet representation of an image to delete as much spatial information as possible while retaining the classifier's prediction. We show theoretically and empirically that our method manages to guard against explanation artifacts, while retaining the ability localize the relevant features remarkably well.

1. Introduction

Modern image classifiers are known to be difficult to interpret and 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 [7, 9, 11, 12, 16, 19, 26]. The explanation mask is optimized to keep only parts of the image that suffice to retain the classification decision. The main advantage of the mask approach over other saliency map methods [6, 32, 36–39, 41, 42] is that the explanation mask is found by optimizing an objective that intuitively aligns with what is defined as relevance. However, Fong and Vedaldi [12] showed that an unregularized explanation mask is very susceptible to explanation artifacts and unreliable. Therefore, current practice [7, 11, 12] heavily regularizes and parameterizes the explanation masks to be smooth. Although smooth explanation masks can communicate useful explanatory information by roughly localizing objects, they are heavily constrained in their ability to delete irrelevant features in the image due to possible smoothness violations. As a result many details such as background elements, texture, and other spatially localized features cannot be deleted although they may be irrelevant to the classifier.

An optimal mask explanation method should be both resistant to artifacts and capable of deleting all irrelevant details. We present a novel method that achieves this by optimizing a sparse mask in the shearlet representation [21] of an image to delete as much spatial information as possible while retaining the classifier's prediction. We theoretically and empirically show that sparsely masking in the shearlet representation instead of the pixel representation guards against artifacts, which we characterize as artificial edges. Due to shearlets' ability to efficiently encode anisotropic features in images, we can localize the relevant image parts extremely well without producing explanation artifacts. Figure 1 illustrates the advantage of ShearletX over smooth pixel mask explanations with an example. For instance, ShearletX for the wolf image shows that the classifier has learned the core characteristics of a wolf: the wolf's

eyes and distinctive snout. Such anisotropic details cannot be efficiently represented by a smooth explanation mask, which picks up details such as teeth and fur to obey the smoothness constraints. In the example, our novel method can tell us that 1) the classifier can correctly predict the wolf from the wolf's eyes and snout shape. 2) Adding the fur and the teeth in the smooth pixel mask explanation double the class score. The examples in Figure 1 further reveal that a good ImageNet classifier only needs a rough sketch of the class instance to correctly classify many images, similar to humans. Such new explanatory insights motivate why ShearletX will be an important new explanation tool for practioners.

2. Related Work

The explainability field has experienced a surge in research activity over the last decade, due to the obvious societal need to explain machine learning models. We focus on explainability aspects of image classifiers, where saliency maps offer an important and useful way of understanding a classifier's prediction. The community has also introduced other tools, such as concept-based methods [18] and inherently interpretable architectures [8]. Nevertheless, saliency map methods comprise the most common tools for explaining image classifiers. In the following, we review current saliency map methods.

Pixel Attribution Methods

Many of the first saliency map methods assign a relevance score to each pixel indicating its relevance for the prediction. Such methods include Gradient Saliency [33], LRP [6], Guided Backprop [37] and Integrated Gradients [38]. Although the fidelity of such explanations can be verified quantitatively, *e.g.* with deletion and insertion curves [6,29], they are to a large degree heuristic and not directly optimizing for a well defined notion of relevance. Moreover, pixel attributions tend to be relatively jittery and have been even shown to be susecptible to adversarial attacks [10]. Other methods, such as LIME [31] and SHAP [25], form coalitions of pixels by first segmenting the image into superpixels and assigning a relevance score to each superpixel. Not only has research exposed various vulnerabilities of LIME and SHAP [35] but superpixels are suboptimal due to their inflexibility. Within a superpixel one cannot say what is relevant and what not.

Pixel Mask Explanations

Mask-based 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 maximizing the probability score for the classification. 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 probability score after masking the image. The explanation mask is found as a solution to the optimization problem

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

where

Wavelet Mask Explanations

3. WaveletX

3.1. Wavelets for Images

3.2. Method

3.3. Theory

4. ShearletX

4.1. Shearlets for Images

4.2. Method

4.3. Theory

5. Experiments

5.1. Artificial Edges and Hallucinations

5.2. Conciseness and Preciseness

6. Limitations

7. Conclusion

8. Introduction

Modern image classifiers are known to be difficult to interpret and 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 [7, 9, 11, 12, 16, 19, 26] frames saliency map computation as an optimization problem over masks. The explanation mask is optimized to keep only parts of the image that suffice to retain the classification decision. The main advantage of the mask approach over other saliency map methods [6, 32, 36–39, 41, 42] is that the explanation mask is found by optimizing an objective that intuitively aligns with what is defined as relevance. Mask explanations need to be robust against corruptions and discriminate between unnecessary and necessary features of the image. Fong and Vedaldi showed that pixel masks are susceptible to corruptions and propose to mitigate this with TV regularization to enforce

smooth masks. However, we identify a problem in TV regularized masks as the TV penalty compromises the ability to discriminate between unnecessary and necessary features of an image, such as texture and shape. A TV regularized mask can be thought of as a window that locates the relevant part fitting the window. However, features inside the window may not be important and difficult to delete due to the TV penalty. To address this, we revisit a recent technique proposed by Kolek et al. that gives explanations by masking in the wavelet domain (WaveletX). We first prove experimentally and theoretically that WaveletX is robust against explanation corruptions, similar to TV regularized pixel masks. Then we show that WaveletX is better suited to discriminate relevant and irrelevant features and thereby avoids the problems of TV regularized pixel masks. Moreover, we identify and address a significant problem of WaveletX. By adding a crucial spatial regularization term, we get rid of blurry backgrounds that are sometimes difficult to interpret, making WaveletX unambiguous and more interpretable. Moreover, we address the fact that Wavelets are not an optimal representation system for piece-wise smooth images and successfully apply the mask explanation approach in the optimal shearlet representation system. Results show that ShearletX can represent edges better in the explanation. Lastly, we also address an overlooked bias of in-distribution perturbations, realized as an inpainting GAN, that were claimed to be most appropriate for mask explanations [7, 16]. An explanation with inpainting GAN perturbations may reflect not what the classifier needs to see for the correct classification but rather what the inpainting GAN needs to infer the class relevant features. Overall, we are the first to identify all three biases of pixel mask explanations: 1) explanation corruption due to artificial edges, 2) inability to discriminate close relevant and irrelevant features due to TV penalty and 3) bias of an inpainting GAN perturbation. Our improved WaveletX and ShearletX are theoretically motivated and well-suited to mitigate all three biases.

9. Related Work

10. Mask Explanations

We model an image as a map $x : \Omega \to [0, 1]$, where Ω is the image domain. For a digital RGB image with $h \times w$ pixels, we have $\Omega = (\{1, \dots, h\} \times \{1, \dots, w\})^3$. For $K \in \mathbb{N}$ labels, we denote the K-simplex, the space of all probability distributions over K labels, as $\Delta^{(K)}$. An image classifier $\Phi : [0, 1]^\Omega \to \Delta^{(K)}$ is a map from image space to the label probability space. We denote a binary mask on an image as a map $m : \Omega \to \{0, 1\}$. Let $u : \Omega \to [0, 1]$ take random values from a probability distribution $u \sim \mathcal{V}$. We adopt the mathematical interpretation of explanation masks as solutions to a rate-distortion optimization problem as in

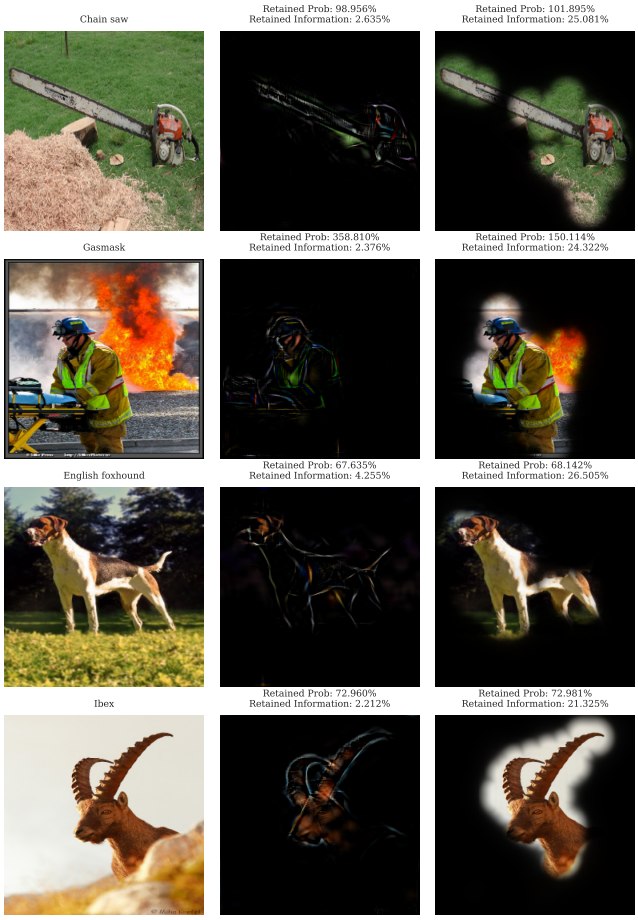


Figure 2. Caption

Kolek et al. in [20]. An optimal explanation mask m^* with sparsity level $s \in \mathbb{N}$ and prior $\mathcal{M} \subset [0, 1]^\Omega$is defined as a solution to

$$\begin{aligned} \min_{m:\Omega\to\{0,1\}} \quad & \mathop{\mathbb{E}}_{u\sim\mathcal{V}} \Big[d(\Phi(x \odot m + (1 - m) \odot u), \Phi(x)) \Big] \\ \text{subject to} \quad & \|m\|_0 \leq s \\ & m \in \mathcal{M}, \end{aligned} \tag{2}$$

where $\odot$ denotes pointwise multiplication and $1 - m$ is pointwise subtraction by abuse of notation. The prior set $\mathcal{M}$ can accomodate certain smoothness conditions such as low total variation and the $d(\cdot, \cdot)$ can be the ℓ_2 norm or set so that the minimization is equivalent to maximization of the predicted label. In practice, the discrete optimization problem is infeasible and the optimal explanation mask is approximated by solving a relaxed optimization problem

$$\min_{m:\Omega\to[0,1]} \quad - \mathop{\mathbb{E}}_{u\sim\mathcal{V}} \Big[\Phi(x \odot m + (1 - m) \odot u) \Big] \tag{3}$$

$$+ \lambda_1 \cdot \|m\|_1 + \lambda_2 \cdot R_{\mathcal{M}}(m), \tag{4}$$

where $R_{\mathcal{M}}(m)$ is typically chosen as a TV penalty to enforce mask smoothness. The optimization is typically performed with projected SGD over the mask m [7, 11, 12, 20]. The perturbation distribution $\mathcal{V}$ has priorly been chosen as Gaussian [12], blurring [12], constant [12], and inpainting GAN [7, 16].

10.1. CartoonX

Kolek et al. [19] applied the mask explanation framework in the wavelet domain of images instead of standard pixel space coined their method *CartoonX*. Since piecewise smooth images are sparsely represented by wavelets [28], sparsely masking in the wavelet domain yields a piecewise smooth (cartoon-like) images. In short, CartoonX was designed to extract the relevant piece-wise smooth part of an image.

Wavelets for Images

A wavelet $\psi : \mathbb{R}^2 \to \mathbb{R}$ is a wave-like oscillation and the cornerstone of the *wavelet transform*

$$\mathcal{W}\{f\}(a,b) = \int_{\mathbb{R}^2} f(u)\frac{1}{a}\psi\Big(\frac{u-b}{a}\Big)\,du \tag{5}$$

of a function $f \in L^2(\mathbb{R}^2)$. The wavelet coefficient $\mathcal{W}\{f\}(a,b)$ extracts time-frequency information of the signal f at position $b \in \mathbb{R}^2$ and scale $a > 0$ (at a frequency proportional to $1/a > 0$). Three suitably chosen mother wavelets $\psi^1, \psi^2, \psi^3 \in L^2(\mathbb{R}^2)$ with dyadic dilations and translations yields an orthonormal basis

$$\Big\{\psi^k_{j,n} := \frac{1}{2^j}\psi^k\Big(\frac{\cdot - 2^j n}{2^j}\Big)\Big\}_{j\in\mathbb{Z}, n\in\mathbb{Z}^2, 1\le k\le 3} \tag{6}$$

of finite energy functions $L^2(\mathbb{R}^2)$. The three indices $k = 1, 2, 3$ correspond to directions (typically vertical, horizontal, and diagonal). A wavelet coefficient $\langle f, \psi^k_{j,n}\rangle$ has high amplitude if the function f has sharp transitions over the support of $\psi^k_{j,n}$. Suitable choices of a (mother) wavelets $\psi^1, \psi^2, \psi^3 \in L^2(\mathbb{R}^2)$, paired with an appropriate scaling function $\phi \in L^2(\mathbb{R}^2)$, define a multiresolution approximation. From the mother wavelet and scaling function one can construct the subspaces

$$\begin{aligned} &V_j := \operatorname{span}\{\phi_{j,n} \mid n \in \mathbb{Z}^2\},\ \phi_{j,n} := 2^{-j/2}\phi(2^{-j}\cdot - n) \\ &W_j := \operatorname{span}\{\psi^k_{j,n} \mid n \in \mathbb{Z}^2, 1 \le k \le 3\}), \end{aligned} \tag{7}$$

which satisfy $V_j \subset V_{j-1}$, $V_j \oplus W_j = V_{j-1}$, and $L^2(\mathbb{R}^2) = \oplus_{j=-\infty}^{J} W_j \oplus V_J$. Therefore, for all $J \in \mathbb{Z}$, any finite energy signal decomposes into

$$f = \sum_{n\in\mathbb{Z}^2} a_n \phi_{J,n} + \sum_{1\le k\le 3}\sum_{j\le J,} d^k_{j,n}\psi^k_{j,n}, \tag{8}$$

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 detail 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 $f \in \text{Ł}^2(\mathbb{R}^2)$ approximated at a certain scale 2^L. The discrete wavelet transform of an image x then computes an invertible wavelet image representation:

$$\mathrm{DWT}(x) = \Big\{a_{J,n}\Big\}_n \cup \Big\{d^1_{j,n}, d^2_{j,n} d^3_{j,n}\Big\}_{L<j\le J,n} \tag{9}$$

corresponding to discretely sampled approximation and detail coefficients of f.

In CartoonX, the explanation mask is optimized on the wavelet image representation. A wavelet mask m is optimized with projected SGD to minimize

$$\begin{aligned} \min_m\ &- \mathop{\mathbb{E}}_{u\sim\mathcal{V}}\Big[\Phi(\mathrm{DWT}^{-1}(m \odot \mathrm{DWT}(x) + (1-m)\odot u))\Big] \\ &+ \lambda_1 \cdot \|m\|_1, \end{aligned} \tag{10}$$

where $\lambda_1 > 0$ is a hyperparameter controling the sparsity of the mask. The sparse wavelet mask is then applied to the wavelet representation and subsequently inverted to yield an approximately piece-wise smooth image $\mathrm{DWT}^{-1}(m \odot \mathrm{DWT}(x))$ that is sufficient to retain the classification decision. Kolek et al. [19] explain that fine details, that are still visible in the CartoonX exlpanation are relevant and fine details that are blurred or blacked out are irrelevant because it is expensive for the wavelet mask to retain fine details and cheap to keep smooth (very blurry) areas. The consequence is that the CartoonX explanation is harder to interpret when smooth image regions are not blacked out. Was the smooth region relevant or was it just inexpensive for the wavelet mask to keep the smooth region (see Figure Fig. 3 for an example)? In the following we propose to modify CartoonX to solve this problem and enforce regions to be black if they are irrelevant. This makes CartoonX easier to interpret and yields a novel kind of explanation method.

Shearlets for Images

The shearlet transform, which was introduced in [14], is based on applying translation, anisotropic dilation, and shearing to generator functions. To dilate and shear 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 $s, a \in Real$. Next, 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

$$\begin{aligned} \psi_{a,s,t,1}(x) &:= a^{-\frac{3}{4}}\psi\left(A_a^{-1}S_s^{-1}(x-t)\right), \\ \psi_{a,s,t,-1}(x) &:= a^{-\frac{3}{4}}\widetilde{\psi}\left(\widetilde{A}_a^{-1}S_s^{-T}(x-t)\right), \end{aligned} \tag{11}$$

where $\widetilde{\psi}(x_1, x_2) := \psi(x_2, x_1)$ for all $x = (x_1, x_2) \in \mathbb{R}^2$. Following [13], we define the continuous shearlet transform as follows:

Definition 10.1 (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 (11) is called a *shearlet system*. The corresponding *(continuous) shearlet transform* is defined by

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

where

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

As we shall see next, if the generator function ψ has directional vanishing moments, then the asymptotic behavior as $a \to 0$ of the continuous shearlet transform of an L^2-function f characterizes its wavefront set. The precise statement can be found in [13].

10.2. CartoonX with Spatial Regularization

To avoid unnecessary blurry regions in CartoonX we penalize unnecessary energy in pixel space with the ℓ_1 norm in pixel space. More precisely, we propose to optimize a wavelet mask m with projected SGD and the optimization objective

$$\min_m \; - \mathop{\mathbb{E}}_{u \sim \mathcal{V}} \Big[\Phi(\mathrm{DWT}^{-1}(m \odot \mathrm{DWT}(x) + (1 - m) \odot u))\Big] + \lambda_1 \cdot \|m\|_1 + \lambda_2 \cdot \|\,\mathrm{DWT}^{-1}(m \odot \mathrm{DWT}(x))\|_1.$$

We illustrate the effect in Figure Fig. 3 by computing CartoonX with and without spatial regularization for five randomly sampled images from Imagenet of different classes. The spatial regularization makes CartoonX much easier to interpret and resolves previous ambiguities.

10.3. CartoonX Theory

Fong and Vedaldi [12] first observed that explanation masks in pixel space are suceptible to explanation artifacts if the mask is not regularized to be smooth. In section Sec. 11.1, we demonstrate that many explanation artifacts of pixel explanation masks stem from artificially formed edges that make up artificial prototypical class features. It is therefore, desirable to have explanations that cannot create artificial edges. In this section, we show theoretically that CartoonX cannot create artificial edges when modeling images and edges in a continuous space. We need to work in continuous space to obtain a rigorous definition of edges but the conclusion is verified later experimentally for discrete images.

In this theory section, we will model images as $L^2([0, 1]^2)$ functions $x : [0, 1]^2 \to [0, 1]$. Edges are singularities, which we will model with the notion of Lipschitz

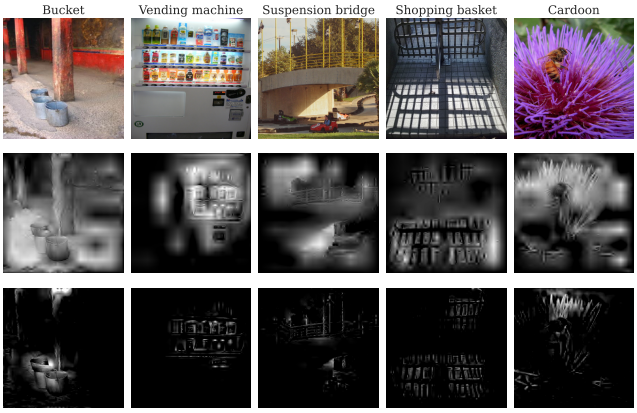


Figure 3. Fair qualitative comparison on images randomly sampled from imagenet. First row: Images from Imagenet with classification label. Second row: CartoonX without spatial regularization from Kolek et al. [19]. Third row: Our CartoonX with spatial regularization.

α regularity, which we introduce below. Our final theorem will show that masking an image $x : [0, 1]^2 \to [0, 1]$ in its wavelet domain cannot not create new (artificial) edges.

Definition 10.2 (Lipschitz α regular functions)**.** A function $f : \mathbb{R}^2 \to \mathbb{R}$ is uniformly Lipschitz $\alpha \geq 0$ over a domain $\Omega \subset \mathbb{R}^2$ if there exists $K > 0$, such that for any $v \in \Omega$ one can find a polynomial p_v of degree $\lfloor \alpha \rfloor$ such that

$$\forall x \in \Omega, \quad |f(x) - p_v(x)| \leq K|x - v|^\alpha. \tag{12}$$

The infimum of K, which satisfies the above equation, is the *homogenous* Hölder α norm $\|f\|_{\dot{C}^\alpha}$. The Hölder α norm of $\|f\|_{C^\alpha} := \|f\|_{\dot{C}^\alpha} + \|f\|_\infty$ also imposes that f is bounded.

For our results, we also need the defintion of *fast decaying derivatives*, which we state below.

Definition 10.3 (Fast Derivative Decay)**.** If $f : [0, 1]^2 \to \mathbb{R}$ is $C^{\lfloor \alpha \rfloor}$ then we say the derivatives of f decay fast if and only if for all multi-indices $\nu = (\nu_1, \nu_2)$ with $\nu_1 + \nu_2 \leq \lfloor \alpha \rfloor$ and for all $m \in \mathbb{N}$ there exists $C_m \in \mathbb{R}_+$ such that

$$\forall u \in [0, 1]^2 : \; |\partial^\nu \psi(t)| \leq \frac{C_m}{1 + |t|^m}. \tag{13}$$

Our result relies on the fact that function regularity as measured by Lipschitz α regularity is characterized by the decay of wavelet coefficients. This is shown in the following theorem, which is a small adaptation of Theorem 9.15 in [27].

Theorem 1. *Let $f : [0, 1]^2 \to \mathbb{R}$ be Lipschitz α on a domain $\Omega_0 \subset [0, 1]^2$. Consider a mother-wavelet ψ with $\lfloor \alpha \rfloor$ vanishing moments. Then, there exists a constant $B > 0$,*

such that for all $1 \leq l \leq 3, j \in \mathbb{Z}, 0 \leq n < 2^{-j}$ *with* $\operatorname{supp} \psi^l_{j,n} \subset \Omega_0$*, we have*

$$|\langle f, \psi^l_{j,n} \rangle| \leq B \cdot \|f\|_{\tilde{C}^\alpha} \cdot 2^{j(\alpha+1)}. \tag{14}$$

Moreover, if ψ *is* $C^{\lfloor \alpha \rfloor}$ *and its derivatives are decaying fast (see Definition 10.3) then there exists a constant* $A > 0$ *such that for all* $1 \leq l \leq 3, j \in \mathbb{Z}, 0 \leq n < 2^{-j}$ *with* $\operatorname{supp} \psi^l_{j,n} \subset \Omega_0$*, we have*

$$A \cdot \|f\|_{\tilde{C}^\alpha} \cdot 2^{j(\alpha+1)} \leq |\langle f, \psi^l_{j,n} \rangle|. \tag{15}$$

In particular, inequality (14) implies f *is Lipschitz* α *on* Ω_0*.*

The Proof is in the supplementary material. The above theorem essentially says that an image's Lipschitz α regularity in a domain $\Omega_0 \subset [0,1]^2$, is completely determined by the decay of its wavelet coefficients, for wavelets $\psi^l_{j,n}$ that have spatial support in Ω_0. Next we define, what we call *the admissible singularity set*, which contains 1d edges as a special case.

Definition 10.4 (Admissible Singularity Set)**.** Let $0 \leq \alpha < 1 \leq \beta$. We say a set $\Gamma \subset [0,1]^2$ is an admissible singularity set of order $(\alpha, \beta) \in \mathbb{R}^2_+$ for an image $x : [0,1]^2 \to \mathbb{R}$ if the following three properties are satisfied:

1. The image x is Lipschitz α on Γ
2. The image x is Lipschitz β on $[0,1]^2 \setminus \Gamma$
3. There exists two constants $0 < K_1 \leq K_2$ such that for all $j \in \mathbb{N}$

$$C_1 2^{-j} \leq \#\{I^{(j)}_{n,m} \mid \Gamma \cap I^{(j)}_{n,m} \neq \emptyset\} \leq C_2 2^{-j}, \tag{16}$$

where $I^{(j)}_{n,m} := [2^{-j} \cdot n, 2^{-j} \cdot (n+1)] \times [2^{-j} \cdot m, 2^{-j} \cdot (m+1)]$.

The most important example of an admissible singularity set $\Gamma \subset [0,1]^2$ of order (α, β) for an image $x : [0,1]^2 \to \mathbb{R}$, is an image that is differentiable outside of Γ (case $\beta = 1$) and has an edge in a 1d curve Γ (case $\alpha = 0$). The property in (16) is satisfied by 1d curves and sets of finitely many points but not by 2d regions. Next, we define the wavelet mask in our coninuous image setting.

Definition 10.5 (Wavelet Mask in Continuum)**.** Consider an image $x : [0,1]^2 \to \mathbb{R}$ and let

$$m_\psi : \{(j,n) \in \mathbb{Z} \times \mathbb{Z}^2 \mid 2^j n \in [0,1)\} \to \{0,1\} \tag{17}$$

be a mask on the wavelet coefficients $\{\langle x, \psi^l_{j,n} \rangle\}_{1 \leq l \leq 3, j \leq J, 2^j n \in [0,1)}$ and let

$$m_\phi : \{n \in \mathbb{Z}^2 \mid 2^j n \in [0,1)\} \to \{0,1\} \tag{18}$$

be a mask on the approximation coefficients $\{\langle x, \phi^l_{J,n} \rangle\}_{2^J n \in [0,1)}$ We define the mean and standard deviation of wavelet coefficients at scale 2^j in x as $\mu_\psi[j]$ and $\sigma_\psi[j]$. They are computed as:

$$\mu_\psi[j] := \frac{1}{3 \cdot 2^{-2j}} \sum_{1 \leq l \leq 3} \sum_{n:\ 2^j n \in [0,1)^2} \langle x, \psi^l_{j,n} \rangle \tag{19}$$

$$\sigma_\psi[j] := \left(\frac{1}{3 \cdot 2^{-2j}} \sum_{1 \leq l \leq 3} \sum_{n:\ 2^j n \in [0,1)^2} |\langle x, \psi^l_{j,n} \rangle - \mu_\psi[j]|^2 \right)^{1/2}$$

Moreover, we define the wavelet obfuscation of x with the wavelet mask m_ψ, m_ϕ and perturbations $\{z_\psi[j,n], z_\phi[J,n]\}_{j \in \mathbb{Z}, n \in \mathbb{Z}^2}$ as

$$\hat{x} = \sum_{1 \leq l \leq 3} \sum_{j \leq J} \sum_{2^j n \in [0,1)^2} \Big(m_\psi[j,n] \langle x, \psi^l_{j,n} \rangle + (1 - m_\psi[j,n]) z_\psi[j,n] \Big) \psi^l_{j,n} \tag{20}$$

$$+ \sum_{2^J n \in [0,1)} \Big(m_\phi[J,n] \langle x, \phi_{J,n} \rangle + (1 - m_\phi[J,n]) z_\phi[J,n] \Big) \phi_{J,n}. \tag{21}$$

Note that the above definition of the wavelet obfuscation $\hat{x}$ only differs from the obfuscation in the CartoonX method (see term in the expectation in Eq. (10)) by a) the fact that images are discrete in practice and by b) masks being only approximately binary. We finally give our final theorem, that shows the obfuscation of $\hat{x}$ with appropriately chosen perturbations cannot create new edges.

Theorem 2 (CartoonX does not create artificial edges.)**.** *Consider an image* $x : [0,1]^2 \to \mathbb{R}$ *in* $L^2([0,1]^2)$ *with an admissible singularity set* $\Gamma \subset [0,1]^2$ *of order* (α, β)*. Let* ψ *be a mother-wavelet with* $\lfloor \alpha \rfloor$ *vanishing moments, derivatives decaying fast and let* ϕ *be a corresponding scaling function for a multiresolution approximation. Then the following holds :*

1. *The mean wavelet coefficient at a fixed scale* 2^j *is zero:*

$$\mu_\psi[j] = 0 \tag{22}$$

and the wavelet coefficient standard deviation at scale 2^j *decays as*

$$\sigma_\psi[j] = \mathcal{O}(2^{j(\min\{\alpha + 3/2, \beta+1\})}). \tag{23}$$

2. *Let* $\mathcal{Q}_\psi \in \{\mathit{Unif}(\mu_\psi[j] - \sigma_\psi[j]2^{j/2}, \mu_\psi[j] + \sigma_\psi[j]2^{j/2}), \mathit{Const}(0)\}$ *and let* $\mathcal{Q}_\phi$ *be an aribitrary distribution over* $\mathbb{R}$*. The obfuscation* $\hat{x}$ *of* x *with any wavelet mask* (m_ψ, m_ϕ) *and perturbations* $z_\psi[j,n] \sim \mathcal{Q}_\psi$ *and* $z_\phi[J,n] \sim \mathcal{Q}_\phi$ *satisfies with probability one*

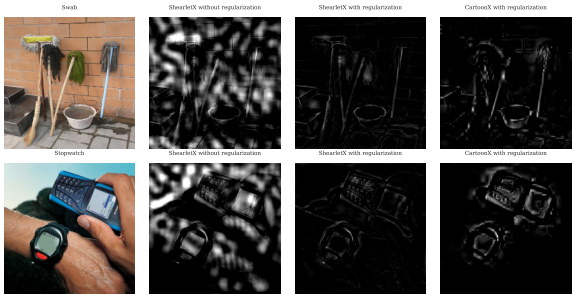


Figure 4. ShearletX

over randomness in $z_\phi[J,n], z_\psi[j,n]$ that $\hat{x}$ is Lipschitz $\alpha+1$ on $[0,1]^2 \setminus \Gamma$, Lipschitz β on Γ, and $\hat{x}$ has no admissible singularity set in $[0,1]^2 \setminus \Gamma$. In partiuclar, $\hat{x}$ cannot contain edges (modeled as Lipschitz $\alpha < 1$ regions) that were not already present in x.

10.4. ShearletX

Wavelets sparsely represent piece-wise smooth (cartoon-like) images but not optimally. In fact, wavelets can only represent point-like singularities optimally. Piece-wise smooth images have edges as singularities and are optimally represented by a different representation system called *shearlets* [22]. Continuous shearlets constitute a 2-parameter dilation group, parametrized by parabolic scaling matrices and shear matrices. The continuous shearlets depend on three parameters, the scaling paramter $a > 0$, the shear parameter $s \in \mathbb{R}$ and the translation parameter $t \in \mathbb{R}^2$, and are defined as

$$\psi_{a,s,t}(u) = a^{-3/4}\psi(D_{a,s}^{-1}(u-t)), \tag{24}$$

where ψ is an appropriately chosen mother wavelet, $u \in \mathbb{R}^2$ and

$$D_{a,s} = \begin{bmatrix} a & -a^{1/2}s \\ 0 & a^{1/2} \end{bmatrix} \tag{25}$$

is a shearing and scaling matrix. The associated continuous shearlet transform is defined by

$$\mathrm{SH}\{f\}(a,s,t) \coloneqq \langle f, \psi_{a,s,t}\rangle. \tag{26}$$

There is of course also a digital shearlet transform, which we use to optimize and explanation mask in the shearlet domain of images. The explanation objective becomes:

$$\min_m \; - \mathop{\mathbb{E}}_{u\sim\mathcal{V}}\Big[\Phi(\mathrm{SH}^{-1}(m\odot \mathrm{SH}(x) + (1-m)\odot u))\Big] + \lambda_1 \cdot \|m\|_1 + \lambda_2 \cdot \|\,\mathrm{SH}^{-1}(m\odot \mathrm{SH}(x))\|_1.$$

We call the resulting explanation method *ShearletX*. In Figure Fig. 4, we compare ShearletX with and without spatial regularization to WaveletX (with spatial regularization). We see that the spatial regularization term that we used to improve WaveletX, is again crucial for ShearletX. ShearletX produces sharper visualizations that are less blurry than WaveletX, which is directly the result of shearlets being better suited to represent piece-wise smooth images sparsely. Although ShearletX produces cleaner visualizations, it is much more computationally expensive due to the shearlet representation having more coefficients. In our implementation, an image of 226×226 pixels has $49 \times 226 \times 226$ shearlet parameters (49 shearing and scaling paramters and $226\times$ locations) unlike the wavelet representation which has 226×226 wavelet coefficients. We leave a more efficient implementation for a faster ShearletX to future work.

10.5. Shearlets and wavefront set

The wavefront set is a concept that characterizes the oriented singularities of distributions, in particular, of $L^2(\mathbb{R}^2)$ functions.

Definition 10.6. [17, 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

$$\big|\widehat{\phi f}(\xi)\big| \le C_k\big(1+|\xi|\big)^{-k} \quad \text{for all } \xi \in \mathbb{R}^2 \setminus \{0\} \text{ s.t. } \xi/|\xi| \in V_\lambda$$

holds for some $C_k > 0$. The *k-wavefront set* $\mathrm{WF}_k(f)$ is the complement of the set of all k-regular directed points and the *wavefront set* $\mathrm{WF}(f)$ is defined as

$$\mathrm{WF}(f) \coloneqq \bigcup_{k\in\mathbb{N}} \mathrm{WF}_k(f),$$

One can make use of the shearlet transform to extract the wavefront set of a given function as follows.

Theorem 3. *Let $f \in L^2(\mathbb{R}^2)$ and assume $(x_0,\lambda_0) \in \mathbb{R}^2 \times \mathbb{S}^1$ is a k-regular directed point of f for some $k \in \mathbb{N}$. Next, consider a continuous shearlet system with generator function $\psi \in H^l(\mathbb{R}^2)$, $l \in \mathbb{N}$, with Fourier transform $\widehat{\psi} \in L^1(\mathbb{R}^2)$ where ψ has $m \in \mathbb{N}$ vanishing moments in x_1-direction, i.e.,*

$$\int_{\mathbb{R}^2} \frac{\big|\widehat{\psi}(\xi_1,\xi_2)\big|^2}{|\xi_1|^{2m}} d\xi < \infty.$$

Finally, assume ψ displays the following asymptotic behavior for $p \in \mathbb{N}$:

$$\big|\psi(x)\big| = \mathcal{O}\big((1+|x|)^{-p}\big) \quad \text{for } |x| \to \infty.$$

Then, there exist a neighborhood $U_0 \subset \mathbb{R}^2$ of x_0 and a neighborhood $S_0 \subset \mathbb{S}^1$ of λ_0 such that

$$\big|\mathcal{SH}_\psi(f)(a,s,x,\iota)\big| = \mathcal{O}\Big(a^{\frac{p}{2}-\frac{3}{4}} + a^{\frac{m}{4}} + a^{\frac{3k}{4}-\frac{3}{4}} + a^{\frac{3l}{4}}\Big) \quad \text{as } a \to 0$$

Method	Time
Integrated Gradients (Sundararajan et al. [38])	0.22s
Smooth Mask (Fong et al. [11])	20s
Wavelet Mask by Kolek et al. [19]	20s
WaveletX (ours)	20s
ShearletX (ours)	20s

Table 1. Computation time for explanation of Resnet18 [15] decision with 256×256 input image. It is well-known that mask explanations are much more computational expensive than more heuristic methods such as Integrated Gradients [38]. Our WaveletX implementation is faster than the implementation of Kolek et al. [19] due to better choices of learning rate. ShearletX is slower than WaveletX due to the mask on the shearlet representation being much larger.

for all $x \in U_0$ *and all* $s \in \mathbb{R}$ *and* $\iota \in \{-1, 1\}$ *such that* $\lambda(s, \iota) \in S_0$, *where*

$$\lambda(s,\iota) := \begin{cases} \left(\frac{1}{\sqrt{s^2+1}}, \frac{s}{\sqrt{s^2+1}} \right) & \text{if } \iota = 1, \\ \left(\frac{s}{\sqrt{s^2+1}}, \frac{1}{\sqrt{s^2+1}} \right) & \text{if } \iota = -1. \end{cases} \tag{27}$$

Remark. Under suitable assumptions on the shearlet generator ψ, the converse of Theorem 3 holds as well. More precisely, following [13], assume that ψ is sufficiently regular for any $k \in \mathbb{N}$. Next, let $(x_0, \lambda_0) \in \mathbb{R}^2 \times \mathbb{S}^1$ and assume there exist a neighborhood $U_0 \subset \mathbb{R}^2$ of x_0 and a neighborhood $S_0 \subset \mathbb{S}^1$ of λ_0 such that

$$\left| \mathcal{SH}_\psi(f)(a, s, x, \iota) \right| = \mathcal{O}(a^n) \quad \text{as } a \to 0,$$

holds for sufficiently large $n \in \mathbb{N}$ uniformly for $x \in U_0$ and all $s \in \mathbb{R}$, $\iota \in \{-1, 1\}$ such that $\lambda(s, \iota) \in S_0$. Then, $(x_0, \lambda_0) \notin \mathrm{WF}_k(f)$.

Theorem 3 and Remark 10.5 demonstrate that the wavefront set is completely determined by the decay properties of the shearlet transform. This implies that in the continuous setting, one can compute the wavefront set of a function by first computing its continuous shearlet transform, then analyzing the pairs of point and direction where this shearlet transform exhibits rapid decay as $a \to 0$.

11. Explanation Bias of Mask Explanations

Mask based explanations can be corrupted by unintended bias. Fong and Vedaldi [12] first showed that pixel space masks without total variation regularization can be corrupt. In light of, our understanding of adversarial vulnerabilities of neural networks, one may think that the explanation corruption comes from uninterpretable adversarial artifacts. However, we empirically find that mostly the corruptions come from artificial edges that make up class prototypes.

11.1. Artificial Edges & Hallucinations

For most instances pixel masks with no or low total variation regularization will highlight a sensible region of the image that activates the class label. When the explanation is corrupted we find it is due to artificial edges forming semantically meaningful concepts that activate the predicted label but do not explain what the classifier saw in the image (see Figure for examples). This can be mitigated with a high TV regularization on the pixel mask. As we show in Section Sec. 11.2, high TV regularization has its own drawback. WaveletX presents itself as a method that is not vulnerable to artificial edges that corrupt the explanation. In Section Sec. 10.3, we proved that CartoonX in continuum cannot produce artificial edges. Our experiments confirm that this translates to the digital case. We do not observe examples as in Figure for WaveletX and we ran experiments that estimate and quantify the amount of artificial edges for pixel masks, WaveletX, each with and without regularization. For that, we compute each explanation on 1000 randomly drawn images from the Imagenet validation set. We quantify the amount of artificial edges in an explanation E for image x with the *hallucination score*:

$$\text{Hallucination Score} := \frac{\#[\text{Edges}(E) \setminus \text{Edges}(x)]}{\#\text{Edges}(x)} \tag{28}$$

The hallucination score counts the number of edges that are present in the explanation E but not in the original image x normalized by the total number of edges in x. To extract the edges we use a shearlet-based edge extractor [5, 30, 40]. To compensate for estimation errors in the edge extractor, we slightly thicken the edges in $\text{Edges}(x)$ for the term $\#[\text{Edges}(E) \setminus \text{Edges}(x)]$. Figure illustrates the edge extractor and artificial edges computed in pixel masks, WaveletX, and ShearletX. In Figure **??**, we present a scatterplot comparing the expected distotion and hallucination score of the 1000 explanations for the distinct mask method. The result shows that pixel masks without TV regularization produce more artificial edges than WaveletX (with and without spatial regularization). In particular, pixel masks without TV regularization have a very fat tail for very high artifact explanations. This reflects what we observed empirically: the majority of explanations are to corrupted but there is a significant amount of outliers that are corrupted by a large amount of artificial edges.

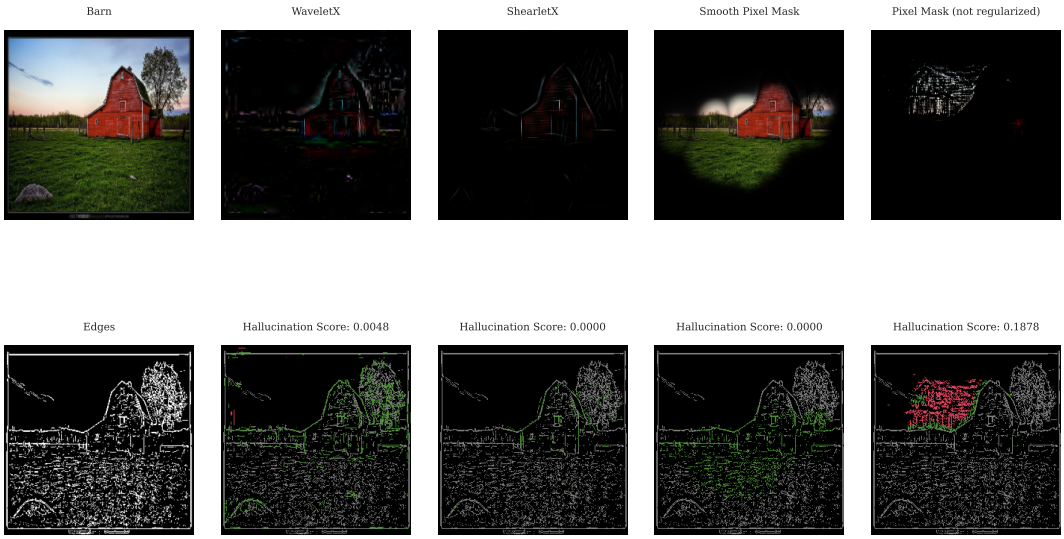


Figure 5. Visualization of Artificial Edges in Explanations

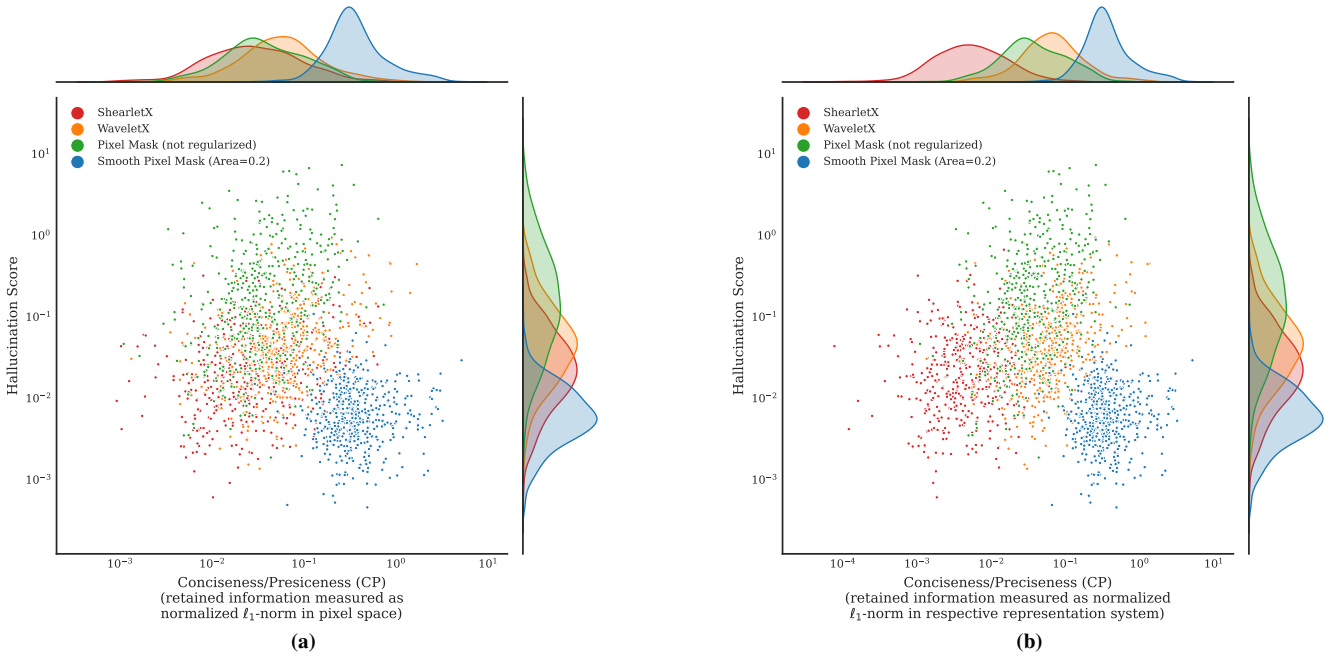


Figure 6. (a) Retained information for conciseness measured as normalized ℓ_1-norm in pixel domain. (b) Retained information for conciseness measured as normalized ℓ_1-norm in respective representation system. Evaluated on Resnet18 [15] on 100 random validation Imagenet samples.

11.2. Discriminating Spatially Close Features

11.3. The Inpainting GAN Bias

12. Sanity Checks

Layer Randomization

Class Localization

Shearletx is not an edge detector

13. Limitations

14. Conclusion

14.1. Language

All manuscripts must be in English.

Method	Retained Probability	Retained Info (Pixel)	Retained Info (Representation)
Smooth Pixel Mask [11] (area = 0.2)	73.76%	%27.55	27.55%
Smooth Pixel Mask [11] (area = 0.1)	51.63%	14.27%	14.72%
Smooth Pixel Mask [11] (area = 0.05)	44.90%	7.41%	7.41%
Pixel Mask (not regularized)	117.66%	3.97%	3.97%
WaveletX (our second best)	66.63%	6.63%	7.21%
ShearletX (our best)	**115.81**%	**2.77**%	**0.74**%

Table 2. Results. Ours is better.

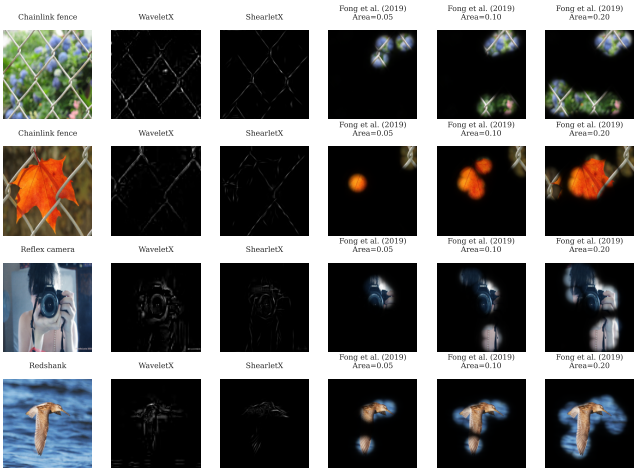


Figure 7. We compare our WaveletX and ShearletX to the smooth masks by Fong et al. (2019) [11] for various area sizes of the mask. Smooth masks fail to unambiguously highlight the chain as relevant.

14.2. Dual submission

Please refer to the author guidelines on the CVPR 2023 web page for a discussion of the policy on dual submissions.

14.3. Paper length

Papers, excluding the references section, must be no longer than eight pages in length. The references section will not be included in the page count, and there is no limit on the length of the references section. For example, a paper of eight pages with two pages of references would have a total length of 10 pages. **There will be no extra page charges for CVPR 2023.**

Overlength papers will simply not be reviewed. This includes papers where the margins and formatting are deemed to have been significantly altered from those laid down by this style guide. Note that this LaTeX guide already sets figure captions and references in a smaller font. The reason such papers will not be reviewed is that there is no provision for supervised revisions of manuscripts. The reviewing process cannot determine the suitability of the paper for presentation in eight pages if it is reviewed in eleven.

14.4. The ruler

The LaTeX style defines a printed ruler which should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document using a non-LaTeX document preparation system, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera-ready copy should not contain a ruler. (LaTeX users may use options of cvpr.sty to switch between different versions.)

Reviewers: note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (*e.g.*, this line is 087.5), although in most cases one would expect that the approximate location will be adequate.

14.5. Paper ID

Make sure that the Paper ID from the submission system is visible in the version submitted for review (replacing the "*****" you see in this document). If you are using the LaTeX template, **make sure to update paper ID in the appropriate place in the tex file**.

14.6. Mathematics

Please number all of your sections and displayed equations as in these examples:

$$E = m \cdot c^2 \tag{29}$$

and

$$v = a \cdot t. \tag{30}$$

It is important for readers to be able to refer to any particular equation. Just because you did not refer to it in the text does not mean some future reader might not need to refer to it. It is cumbersome to have to use circumlocutions like "the equation second from the top of page 3 column 1". (Note that the ruler will not be present in the final copy, so is not

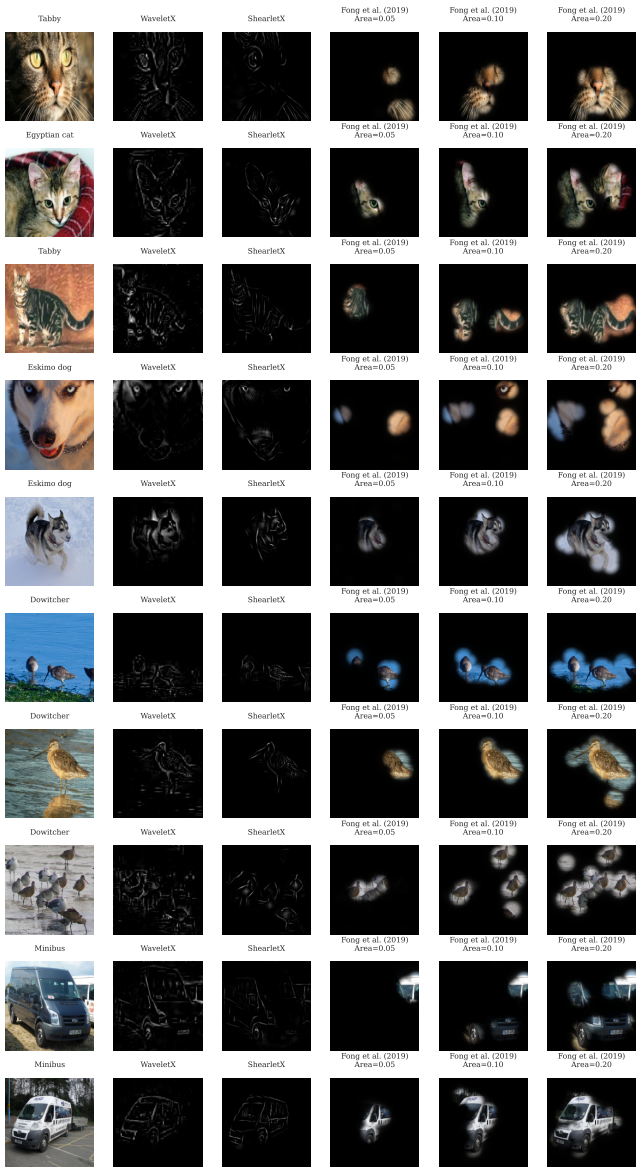


Figure 8. We compare our WaveletX and ShearletX to the smooth masks by Fong et al. (2019) [11] for various area sizes of the mask. Smooth masks fail to unambiguously highlight the chain as relevant.

an alternative to equation numbers). All authors will benefit from reading Mermin's description of how to write mathematics: http://www.pamitc.org/documents/mermin.pdf.

14.7. Blind review

Many authors misunderstand the concept of anonymizing for blind review. Blind review does not mean that one must remove citations to one's own work—in fact it is often impossible to review a paper unless the previous citations are known and available.

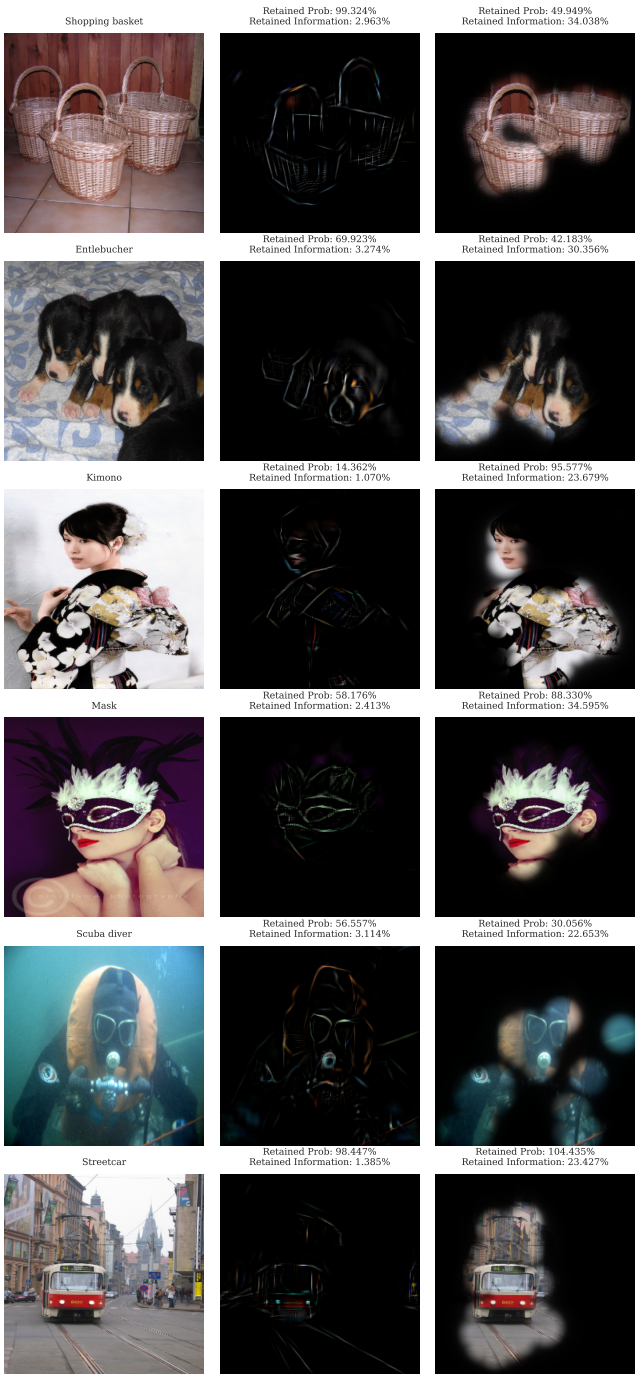


Figure 9. Caption

Blind review means that you do not use the words "my" or "our" when citing previous work. That is all. (But see below for tech reports.)

Saying "this builds on the work of Lucy Smith [1]" does not say that you are Lucy Smith; it says that you are building on her work. If you are Smith and Jones, do not say "as we show in [7]", say "as Smith and Jones show in [7]" and at

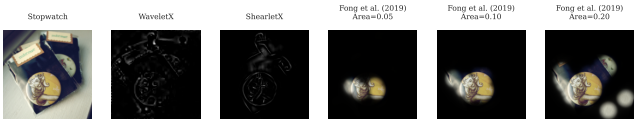


Figure 10. Comparing WaveletX and ShearletX to the smooth masks by Fong et al. (2019) [11] for various area sizes of the mask on various misclassified images.

the end of the paper, include reference 7 as you would any other cited work.

An example of a bad paper just asking to be rejected:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of our previous paper [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Removed for blind review

An example of an acceptable paper:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of the paper of Smith *et al*. [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Smith, L and Jones, C. "The frobnicatable foo filter, a fundamental contribution to human knowledge". Nature 381(12), 1-213.

If you are making a submission to another conference at the same time, which covers similar or overlapping material, you may need to refer to that submission in order to explain the differences, just as you would if you had previously published related work. In such cases, include the anonymized parallel submission [23] as supplemental material and cite it as

> [1] Authors. "The frobnicatable foo filter", F&G 2014 Submission ID 324, Supplied as supplemental material `fg324.pdf`.

Finally, you may feel you need to tell the reader that more details can be found elsewhere, and refer them to a technical report. For conference submissions, the paper must stand on its own, and not *require* the reviewer to go to a tech report for further details. Thus, you may say in the body of the paper "further details may be found in [24]". Then submit the tech report as supplemental material. Again, you may not assume the reviewers will read this material.

Sometimes your paper is about a problem which you tested using a tool that is widely known to be restricted to a single institution. For example, let's say it's 1969, you have solved a key problem on the Apollo lander, and you believe that the CVPR70 audience would like to hear about your solution. The work is a development of your celebrated 1968 paper entitled "Zero-g frobnication: How being the only people in the world with access to the Apollo lander source code makes us a wow at parties", by Zeus *et al*.

You can handle this paper like any other. Do not write "We show how to improve our previous work [Anonymous, 1968]. This time we tested the algorithm on a lunar lander [name of lander removed for blind review]". That would be silly, and would immediately identify the authors. Instead write the following:

> We describe a system for zero-g frobnication. This system is new because it handles the following cases: A, B. Previous systems [Zeus et al. 1968] did not handle case B properly. Ours handles it by including a foo term in the bar integral.
>
> ...
>
> The proposed system was integrated with the Apollo lunar lander, and went all the way to the moon, don't you know. It displayed the following behaviours, which show how well we solved cases A and B: ...

As you can see, the above text follows standard scientific convention, reads better than the first version, and does not explicitly name you as the authors. A reviewer might think it likely that the new paper was written by Zeus *et al*., but cannot make any decision based on that guess. He or she would have to be sure that no other authors could have been contracted to solve problem B.

FAQ

Q: Are acknowledgements OK?
A: No. Leave them for the final copy.

Q: How do I cite my results reported in open challenges?
A: To conform with the double-blind review policy, you can report results of other challenge participants together with your results in your paper. For your results, however, you should not identify yourself and should not mention your participation in the challenge. Instead present your results referring to the method proposed in your paper and draw conclusions based on the experimental comparison to other results.

14.8. Miscellaneous

Compare the following:

`$conf_a$`	$conf_a$
`$\mathit{conf}_a$`	conf_a

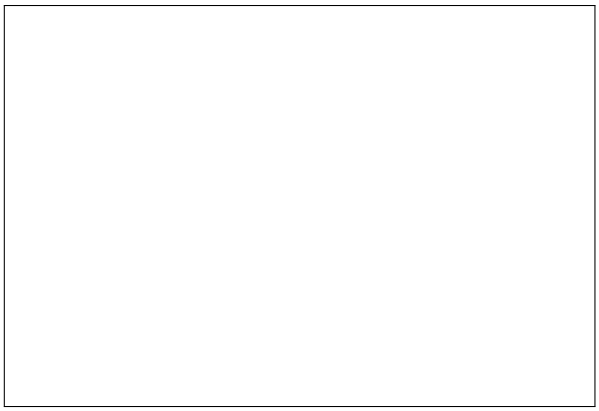


Figure 11. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

See The TEXbook, p165.

The space after *e.g.*, meaning "for example", should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al.*" (not "*et. al.*" as "*et*" is a complete word). If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al.* However, use it only when there are three or more authors. Thus, the following is correct: "Frobnication has been trendy lately. It was introduced by Alpher [1], and subsequently developed by Alpher and Fotheringham-Smythe [2], and Alpher *et al.* [3]."

This is incorrect: "... subsequently developed by Alpher *et al.* [2] ..." because reference [2] has just two authors.

15. Formatting your paper

All text must be in a two-column format. The total allowable size of the text area is $6\frac{7}{8}$ inches (17.46 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be $1\frac{1}{8}$ inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately $1\frac{5}{8}$ inches (4.13 cm) from the bottom edge of the page.

15.1. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area $6\frac{7}{8}$ inches (17.46 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Page numbers should be in the footer, centered and $\frac{3}{4}$ inches from the bottom of the page. The review version should have page numbers, yet the final version submitted as camera ready should not show any page numbers. The LaTeX template takes care of this when used properly.

15.2. Type style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title $1\frac{3}{8}$ inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. $\frac{1}{6}$ inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Figs. 11 and 12. Short captions should be centred.

Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

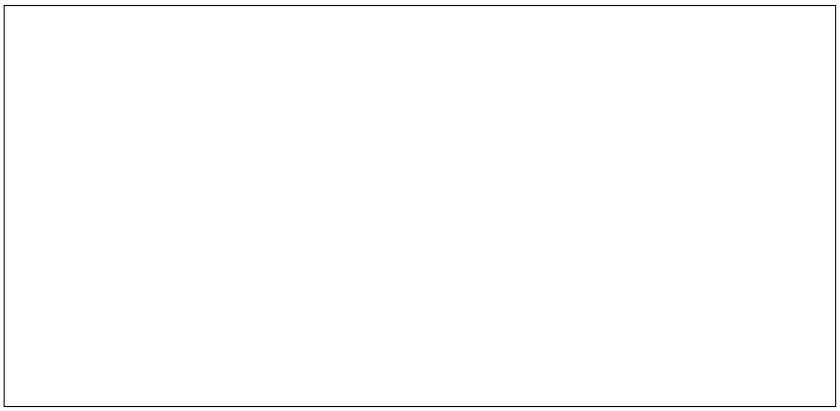
FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

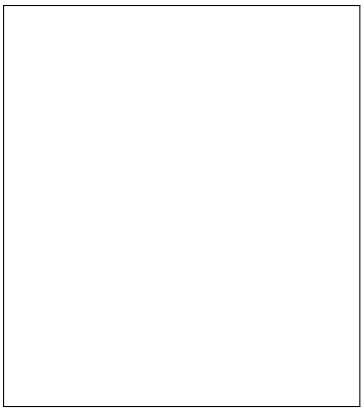
15.3. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the

[1]This is what a footnote looks like. It often distracts the reader from the main flow of the argument.



(a) An example of a subfigure.



(b) Another example of a subfigure.

Figure 12. Example of a short caption, which should be centered.

bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

15.4. Cross-references

For the benefit of author(s) and readers, please use the

```
\cref{...}
```

command for cross-referencing to figures, tables, equations, or sections. This will automatically insert the appropriate label alongside the cross-reference as in this example:

> To see how our method outperforms previous work, please see Fig. 11 and Tab. 3. It is also possible to refer to multiple targets as once, *e.g.* to Figs. 11 and 12a. You may also return to Sec. 15 or look at Eq. (30).

If you do not wish to abbreviate the label, for example at the beginning of the sentence, you can use the

```
\Cref{...}
```

command. Here is an example:

> Figure 11 is also quite important.

15.5. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [23]. Where appropriate, include page numbers and the name(s) of editors of referenced books. When you cite multiple papers at once, please make sure that you cite them in numerical order like this [1, 2, 4, 23, 24]. If you use the template as advised, this will be taken care of automatically.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 3. Results. Ours is better.

15.6. Illustrations, graphs, and photographs

All graphics should be centered. In LaTeX, avoid using the `center` environment for this purpose, as this adds potentially unwanted whitespace. Instead use

```
\centering
```

at the beginning of your figure. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths that render effectively in print. Readers (and reviewers), even of an electronic copy, may choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.pdf}
```

15.7. Color

Please refer to the author guidelines on the CVPR 2023 web page for a discussion of the use of color in your document.

If you use color in your plots, please keep in mind that a significant subset of reviewers and readers may have a color

vision deficiency; red-green blindness is the most frequent kind. Hence avoid relying only on color as the discriminative feature in plots (such as red *vs*. green lines), but add a second discriminative feature to ease disambiguation.

16. Final copy

You must include your signed IEEE copyright release form when you submit your finished paper. We MUST have this form before your paper can be published in the proceedings.

Please direct any questions to the production editor in charge of these proceedings at the IEEE Computer Society Press: https://www.computer.org/about/contact.

References

[1] FirstName Alpher. Frobnication. *IEEE TPAMI*, 12(1):234–778, 2002. 13, 14

[2] FirstName Alpher and FirstName Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003. 13, 14

[3] FirstName Alpher, FirstName Fotheringham-Smythe, and FirstName Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004. 13

[4] FirstName Alpher and FirstName Gamow. Can a computer frobnicate? In *CVPR*, pages 234–778, 2005. 14

[5] Andrade-Loarca, Kutyniok, and Öktem. Shearlets as feature extractor for semantic edge detection: the model-based and data-driven realm. *R. Soc. A*, Volume 476, Issue 2243, 2020. 8

[6] 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. 1, 2

[7] 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, 2, 3, 4

[8] 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

[9] 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, 2

[10] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[11] 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, 2, 4, 8, 10, 11, 12

[12] 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, 4, 5, 8

[13] Philipp Grohs. Continuous shearlet frames and resolution of the wavefront set. *Monatsh. Math.*, 164(4):393–426, 2011. 5, 8

[14] 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,. 4

[15] 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. 8, 9

[16] Cosmas Heiß, Ron Levie, Cinjon Resnick, Gitta Kutyniok, and Joan Bruna. In-distribution interpretability for challenging modalities. *Preprint arXiv:2007.00758*, 2020. 1, 2, 3, 4

[17] Lars Hormander. *The Analysis of Linear Partial Differential Operators. I, Distribution Theory and Fourier Analysis*. Grundlehren Der Mathematischen Wissenschaften. Springer, 1990. 7

[18] 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

[19] 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, 2, 4, 5, 8

[20] Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. A rate-distortion framework for explaining black-box model decisions. In *xxAI - Beyond Explainable AI*, page 91–115, 2022. 3, 4

[21] G. Kutyniok and D. Labate. *Shearlets: Multiscale Analysis for Multivariate Data*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 2012. 1

[22] Gitta Kutyniok and Wang-Q Lim. Compactly supported shearlets are optimally sparse. *Journal of Approximation Theory*, 163(11):1564–1589, 2011. 7

[23] FirstName LastName. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as supplemental material `fg324.pdf`. 12, 14

[24] FirstName LastName. Frobnication tutorial, 2014. Supplied as supplemental material `tr.pdf`. 12, 14

[25] 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

[26] Jan Macdonald, Stephan Wäldchen, Sascha Hauch, and Gitta Kutyniok. A rate-distortion framework for explaining neural network decisions. *Preprint arXiv:1905.11092*, 2019. 1, 2

[27] Stéphane Mallat. *A wavelet tour of signal processing (2. ed.)*. Academic Press, 1999. 5

[28] Stéphane Mallat. *A Wavelet Tour of Signal Processing (Third Edition)*, chapter 11.3. Academic Press, third edition edition, 2009. 4

[29] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018. 2

[30] Rafael Reisenhofer, Johannes Kiefer, and Emily King. Shearlet-based detection of flame fronts. *Experiments in Fluids*, 57, 02 2016. 8

[31] 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

[32] 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. 1, 2

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

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 1

[35] 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

[36] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *Workshop on Visualization for Deep Learning, ICML*, 2017. 1, 2

[37] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. 1, 2

[38] 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. 1, 2, 8

[39] Jacopo Teneggi, Alexandre Luster, and Jeremias Sulam. Fast hierarchical games for image explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–11, 2022. 1, 2

[40] Sheng Yi, Demetrio Labate, Glenn R. Easley, and Hamid Krim. A shearlet approach to edge analysis and detection. *IEEE Transactions on Image Processing*, 18(5):929–941, 2009. 8

[41] Matthew D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1, 2

[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 1, 2

The proof of Theorem