A. Theory

In this section, we give the supplementary material for our theoretical results. We formally proof Theorem 1 and Theorem 2 from the main paper that show masking in the shearlet or wavelet space cannot create artificial edges for continuous images.

A.1. ShearletX Theoretical Result

We begin by recounting the definition of the wavefront set, which is a good model edges for continuous images, particularly when working with shearlets.

Definition A.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

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

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

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

The wavefront set is completely determined by the decay properties of the shearlet transform, which is formalized in the following Theorem from [26].

Theorem 3 (Theorem 2 in [26])**.** *Let ψ be a Schwartz function with infinitely many vanishing moments in x_1-direction. Let f be a tempered distribution and $D = D_1 \cup D_2$, where*

$$\begin{aligned} D_1 = \{&(t_0, s_0) \in \mathbb{R}^2 \times [-1, 1] : \textit{for} \\ &(s, t) \textit{ in a neighborhood } U \textit{ of } (s_0, t_0), \\ &|\mathcal{SH}_\psi(f)(a, s, t)| = O(a^k), \textit{for all } k \in \mathbb{N}, \\ &\textit{with the implied constant uniform over } U)\} \end{aligned}$$

and

$$\begin{aligned} D_1 = \{&(t_0, s_0) \in \mathbb{R}^2 \times (1, \infty] : \textit{for} \\ &(1/s, t) \textit{ in a neighborhood } U \textit{ of } (s_0, t_0), \\ &|\mathcal{SH}_\psi(f)(a, s, t)| = O(a^k), \textit{for all } k \in \mathbb{N}, \\ &\textit{with the implied constant uniform over } U)\}. \end{aligned}$$

Then

$$WF(f)^c = D. \tag{13}$$

For the following theorem, we model the edges in the image x by the wavefront set $\mathrm{WF}(x)$.

Theorem 4. *Let $x \in L^2[0,1]^2$ be an image modeled as a L^2-function. Let m be a mask on the shearlet coefficients of x and let $\hat{x}$ be the image x masked in shearlet space with m. Then, we have $WF(\hat{x}) \subset WF(x)$ and thus masking in shearlet space did not create new edges.*

Proof. Note that the shearlet transform is invertible. Hence, we have by definition of $\hat{x}$

$$\mathcal{SH}(\hat{x})(a, s, t) = \mathcal{SH}(x)(a, s, t) \cdot m(a, s, t). \tag{14}$$

To show $\mathrm{WF}(\hat{x}) \subset \mathrm{WF}(x)$, it suffices to show $\mathrm{WF}^c(\hat{x}) \supset \mathrm{WF}^c(x)$. Let $(t, s) \in \mathrm{WF}^c(x)$ be arbitrary with $|s| < 1$. Then, by definition of the wavefront set, we have for all $N \in \mathbb{N}$

$$|\mathcal{SH}(x)(a, s, t)| = O(a^N) \tag{15}$$

for $a \to 0$. Since $m(a, s, t) \in [0, 1]$, we also have for all $N \in \mathbb{N}$

$$\begin{aligned} |\mathcal{SH}(\hat{x})(a, s, t)| &= |\mathcal{SH}(x)(a, s, t)| \cdot |m(a, s, t)| & (16) \\ &\le |\mathcal{SH}(x)(a, s, t)| = O(a^N). & (17) \end{aligned}$$

This implies $(t, s) \in \mathrm{WF}^c(\hat{x})$. Thus, we showed the claim $\mathrm{WF}^c(\hat{x}) \supset \mathrm{WF}^c(x)$. □

A.2. WaveletX Theoretical Result

When analyzing WaveletX, we opt to model singularities via local Lipschitz regularity (see Definition A.2) 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 (see Theorem 5).

Definition A.2 (Lipschitz Regularity)**.** A function $f : \mathbb{R}^2 \to \mathbb{R}$ is *uniformly Lipschitz* $\alpha \ge 0$ 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

$$\forall x \in \Omega, \ \ |f(x) - p_v(x)| \le K|x - v|^\alpha. \tag{18}$$

The *Lipschitz regularity* of f over Ω is the supermum over all α, for which f is uniformly Lipschitz α over Ω. The infimum of K, which satisfies the above equation, is the *homogenous Hölder* α norm $\|f\|_{\tilde{C}^\alpha}$.

Theorem 5 (Theorem 9.15 [31])**.** *Let $x \in L^2[0,1]^2$ be a continuous image with Lipschitz regularity $\alpha \ge 0$. Then there exist $B \ge A > 0$ such that for all $J \in \mathbb{Z}$ we have*

$$A\|x\|_{\tilde{C}^\alpha} \le \sup_{1 \le l \le 3, j \le J, 2^j n \in [0,1)^2} \frac{|\langle x, \psi^l_{j,n}\rangle|}{2^{j(\alpha+1)}} \le B\|x\|_{\tilde{C}^\alpha}.$$

In our Theorem 5, we will heavily rely on the connection between Lipschitz regularity and wavelet decay that is formalized in Theorem 5. As preparation for our result, we first give a corollary to Theorem 5 that shows a function is uniformly Lipschitz α if and only if the wavelet coefficients decay faster than $\mathcal{O}(2^{j(\alpha+1)})$ for $j \to -\infty$.

Corollary 5.1. *Let $a, b \in \mathbb{R}$ with $a < b$ and consider a continuous image $x \in L^2[a,b]^2$ with square domain $[a,b]^2$. Then the following two statements are equivalent:*

1. *The function x is uniformly Lipschitz α.*
2. *There exists a constant $C > 0$ such that for all $J \in \mathbb{Z}$ with $J \leq 0$, we have*

$$\sup_{1 \leq l \leq 3, j \leq J, 2^j n \in [a,b)^2} \frac{|\langle x, \psi^l_{j,n}\rangle|}{2^{j(\alpha+1)}} \leq C.$$

Proof. We prove the Corollary for $[a,b]^2 = [0,1]^2$ and the general case follows simply with a scaling argument. First, we prove (1) implies (2). Suppose the function $x \in L^2[0,1]^2$ is uniformly Lipschitz α. Then x has Lipschitz regularity $\alpha^* \geq \alpha$. By Theorem 5, we then obtain a constant $\tilde{B} > 0$ such that for all $J \in Z$

$$\sup_{1 \leq l \leq 3, j \leq J, 2^j n \in [0,1)^2} \frac{|\langle x, \psi^l_{j,n}\rangle|}{2^{j(\alpha^*+1)}} \leq \tilde{B}\|x\|_{\tilde{C}^{\alpha^*}}.$$

We then have

$$\begin{aligned}
&\sup_{1 \leq l \leq 3, j \leq J, 2^j n \in [0,1)^2} \frac{|\langle x, \psi^l_{j,n}\rangle|}{2^{j(\alpha+1)}} \\
&= \sup_{1 \leq l \leq 3, j \leq J, 2^j n \in [0,1)^2} \frac{|\langle x, \psi^l_{j,n}\rangle|}{2^{j(\alpha^*+1)}} 2^{j(\alpha^*-\alpha)} \\
&\leq \tilde{B}\|x\|_{\tilde{C}^{\alpha^*}} \sup_{j \leq J} 2^{j(\alpha^*-\alpha)} \\
&= \tilde{B}\|x\|_{\tilde{C}^{\alpha^*}} 2^{J(\alpha^*-\alpha)} \\
&\leq \tilde{B}\|x\|_{\tilde{C}^{\alpha^*}}.
\end{aligned}$$

By setting $C := \tilde{B}\|x\|_{\tilde{C}^{\alpha^*}}$, we have shown (1) implies (2). Next, we show that (2) implies (1). Suppose there exists a constant C such that for all $J \in \mathbb{Z}$ with $J < 0$, we have

$$\sup_{1 \leq l \leq 3, j \leq J, 2^j n \in [a,b)^2} \frac{|\langle x, \psi^l_{j,n}\rangle|}{2^{j(\alpha+1)}} \leq C. \tag{19}$$

We prove x is uniformly Lipschitz α by contradiction. Supposed x has Lipschitz regularity β with $0 \leq \beta < \alpha$. We then have by Theorem 5 that there exists constants $A > 0$ such that for all $J \in \mathbb{Z}$

$$A\|x\|_{\tilde{C}^\beta} \leq \sup_{1 \leq l \leq 3, j \leq J, 2^j n \in [0,1)^2} \frac{|\langle x, \psi^l_{j,n}\rangle|}{2^{j(\beta+1)}}. \tag{20}$$

By taking $J \to -\infty$ in (20), we obtain a sequence of $(j_k)_{k \in \mathbb{N}}$ with $j_k \to -\infty$ satisfying

$$\frac{A}{2}\|x\|_{\tilde{C}^\beta} 2^{j_k(\beta+1)} \leq |\langle x, \psi^l_{j_k,n}\rangle| \leq C 2^{j_k(\alpha+1)}, \tag{21}$$

for all $k \in \mathbb{N}$. But this is a contradiction, since for large enough $k \in \mathbb{N}$, j_k is so negative that the upper bound in (21) is strictly smaller then the lower bound in (21). Thus, x must be uniformly Lipschitz α, which finishes the proof. □

Next, we define a Lipschitz regular point of an image as a point for which the image has locally Lipschitz regularity α with $\alpha \geq 1$.

Definition A.3 (Lipschitz Regular Point)**.** Let $x \in L^2[0,1]^2$ be a continuous image. Let $g : \mathbb{R} \to \mathbb{R}$ be a smooth cutoff function satisfying the following properties:

1. $\forall t \in \mathbb{R} : \; |t| \leq 1/2 \Longrightarrow g(t) = 1$
2. $\forall t \in \mathbb{R} : \; |t| \geq 1 \Longrightarrow g(t) = 0$
3. $\forall t \in \mathbb{R} : \; |g(t)| \leq 1$

Define the 2d cutoff function $h : \mathbb{R}^2 \to \mathbb{R}$, $h(t_1, t_2) := g(t_1)g(t_2)$. We say a point $t^* = (t_1^*, t_2^*) \in [0,1]^2$ is a regular point of x if there exists $0 < a \leq 1$ such that the localized image $\tilde{x} : [0,1]^2 \to \mathbb{R}$,

$$\tilde{x}(t_1, t_2) := h((t_1 - t_1^*)/a, (t_2 - t_2^*)/a) \cdot x(t_1, t_2)$$

has Lipschitz regularity $\alpha \geq 1$.

A Lipschitz singular point is any point that is not a Lipschitz regular point. Lipschitz singular points model image elements such as edges and point singularities.

Theorem 6. *Let $x \in L^2[0,1]^2$ be an image. Consider an orthonormal wavelet basis that comprises compactly supported wavelets. Let m be a bounded mask in wavelet space and denote by $\hat{x}$ the image x masked in wavelet space with m. Then, every Lipschitz regular point t^* of x is also a Lipschitz regular point of $\hat{x}$.*

Proof. Let $t^* = (t_1^*, t_2^*) \in [0,1]^2$ be a Lipschitz regular point of x. By definition, there exists $0 < a \leq 1$, such that the localized image

$$\tilde{x} : [0,1]^2 \to \mathbb{R}, \tag{22}$$
$$\tilde{x}(t_1, t_2) := h((t_1 - t_1^*)/a, (t_2 - t_2^*)/a) \cdot x(t_1, t_2) \tag{23}$$

has Lipschitz regularity $\alpha \geq 1$, where $h : \mathbb{R}^2 \to \mathbb{R}$ is the smooth cutoff function from Definition A.3). By Theorem 5, there exists a constant $B > 0$ such that for every $J \in \mathbb{Z}$

$$\sup_{1 \leq l \leq 3, j \leq J, 2^j n \in [0,1)^2} \frac{|\langle \tilde{x}, \psi^l_{j,n}\rangle|}{2^{j(\alpha+1)}} \leq B\|\tilde{x}\|_{\tilde{C}^\alpha}. \tag{24}$$

By definition of the smooth cutoff function h, we know that h is equal to 1 on the square $S_a(t^*)$ with side length a, centered at $t^* \in [0,1]^2$. For each $j \in \mathbb{Z}$, we define the set

$$\Omega_j := \left\{(n_1, n_2) \in \mathbb{N}^2 : \operatorname{supp} \psi_{j,n} \subset S_a(t^*)\right\}$$

of grid locations at scale 2^j where the wavelet support is contained in $S_a(t^*)$. Note there exists a sufficiently small $J^* \in \mathbb{Z}$, such that $\Omega_j \neq \emptyset$, for all $j \leq J_0$. Fix such a $J^* \in \mathbb{Z}$. Then, for all $j \leq J^*$ and $n \in \Omega_j$, we have

$$\forall t \in \operatorname{supp} \psi^l_{j,n} : \ \tilde{x}(t) = x(t) \tag{25}$$

because $\tilde{x}$ is obtained as x times a cutoff function, which is equal to 1 on the square $S_a(t^*) \supset \operatorname{supp} \psi^l_{j,n} \ni t$. Since the wavelet system is assumed to be an orthonormal basis, the wavelet coefficients of $\hat{x}$ are equal to the masked wavelet coefficients of x, namely $\langle \hat{x}, \psi^l_{j,n} \rangle = m_{j,n} \langle x, \psi^l_{j,n} \rangle$, where $m_{j,n}$ is the mask entry for the wavelet coefficient with parameters (j, n). We then have

$$\begin{aligned}
&\sup_{1 \leq l \leq 3, j \leq J^*, n \in \Omega_j} \frac{|\langle \hat{x}, \psi^l_{j,n} \rangle|}{2^{j(\alpha+1)}} \\
&= \sup_{1 \leq l \leq 3, j \leq J^*, n \in \Omega_j} \frac{|m_{j,n}| |\langle x, \psi^l_{j,n} \rangle|}{2^{j(\alpha+1)}}.
\end{aligned}$$

Without loss of generality we can assume $\sup_{j,n} |m_{j,n}| \leq 1$, otherwise we would just add a constant factor to the analysis. We obtain

$$\begin{aligned}
&\sup_{1 \leq l \leq 3, j \leq J^*, n \in \Omega_j} \frac{|\langle \hat{x}, \psi^l_{j,n} \rangle|}{2^{j(\alpha+1)}} && (26) \\
&\leq \sup_{1 \leq l \leq 3, j \leq J^*, n \in \Omega_j} \frac{|\langle x, \psi^l_{j,n} \rangle|}{2^{j(\alpha+1)}} && (27) \\
&= \sup_{1 \leq l \leq 3, j \leq J^*, n \in \Omega_j} \frac{|\langle \tilde{x}, \psi^l_{j,n} \rangle|}{2^{j(\alpha+1)}}, && (28)
\end{aligned}$$

where we used property (25) for the last equality. We further upper bound the expression by taking the supremum over a larger set of indices:

$$\begin{aligned}
&\sup_{1 \leq l \leq 3, j \leq J^*, n \in \Omega_j} \frac{|\langle \tilde{x}, \psi^l_{j,n} \rangle|}{2^{j(\alpha+1)}} && (29) \\
&\leq \sup_{1 \leq l \leq 3, j \leq J^*, 2^j n \in [0,1)^2} \frac{|\langle \tilde{x}, \psi^l_{j,n} \rangle|}{2^{j(\alpha+1)}} \leq B \|\tilde{x}\|_{\tilde{C}^\alpha}, && (30)
\end{aligned}$$

where we used the upper bound on the wavelet decay from (24) for the last inequality. Overall, we showed

$$\sup_{1 \leq l \leq 3, j \leq J^*, n \in \Omega_j} \frac{|\langle \hat{x}, \psi^l_{j,n} \rangle|}{2^{j(\alpha+1)}} \leq B \|\tilde{x}\|_{\tilde{C}^\alpha}. \tag{31}$$

Next, choose $a' := a/2$ and consider the smaller square $S_{a'}(t^*) \subset S_a(t^*)$ of side length a' which is centered at t^*. There exists a sufficiently small scale $J_0 \in \mathbb{Z}$ so that wavelets with scale parameter $j \leq J_0$ whose support intersects $S_{a'}(t^*)$ must be contained in $S_a(t^*)$. Namely, for all $j \leq J_0$ and $n \in \mathbb{Z}^2$, we have

$$\operatorname{supp} \psi_{j,n} \cap S_{a'}(t^*) \neq \emptyset \implies \operatorname{supp} \psi_{j,n} \subset S_a(t^*). \tag{32}$$

Next, we project $\hat{x}$ to have only scales smaller than 2^{J_0} with the projection operator P_{J_0}:

$$P_{J_0} \hat{x} := \sum_{j \leq J_0} \sum_{2^j n \in [0,1)^2} \langle \hat{x}, \psi^l_{j,n} \rangle \psi^l_{j,n}. \tag{33}$$

We show next that $P_{J_0} \hat{x}$ is uniformly Lipschitz $\alpha \geq 1$ on $S_{a'}(t^*)$. We have, for every $J \in \mathbb{Z}$,

$$\begin{aligned}
&\sup_{1 \leq l \leq 3, j \leq J, 2^j n \in S_a(t^*)} \frac{|\langle P_{J_0} \hat{x}, \psi^l_{j,n} \rangle|}{2^{j(\alpha+1)}} && (34) \\
&\leq \sup_{1 \leq l \leq 3, j \leq J_0, n \in \Omega_j} \frac{|\langle P_{J_0} \hat{x}, \psi^l_{j,n} \rangle|}{2^{j(\alpha+1)}} && (35) \\
&= \sup_{1 \leq l \leq 3, j \leq J_0, n \in \Omega_j} \frac{|\langle \hat{x}, \psi^l_{j,n} \rangle|}{2^{j(\alpha+1)}} && (36) \\
&\leq B \|\tilde{x}\|_{\tilde{C}^\alpha}, && (37)
\end{aligned}$$

where we used in the equality (35) that the wavelet coefficients of $P_{J_0} \hat{x}$ for scale parameters $j > J_0$ are zero and equation (32). In equality (36), we used that $\langle P_{J_0} \hat{x}, \psi^l_{j,n} \rangle = \langle \hat{x}, \psi^l_{j,n} \rangle$ for all $j \leq J_0$, and for the last inequality (37) we used the inequality in (31) where J^* can be chosen as J_0. We can apply now Corollary 5.1 to $P_{J_0} \hat{x}$, which shows that $P_{J_0} \hat{x}$ is uniformly Lipschitz $\alpha \geq 1$ on the domain $S_{a'}(t^*)$. The Lipschitz α property is determined by the asymptotics of the wavelet coefficients for scales going to 0. Therefore, if the projection $P_{J_0} \hat{x}$ is uniformly Lipschitz α on the domain $S_{a'}(t^*)$ then so is $\hat{x}$ uniformly Lipschitz α on the domain $S_{a'}(t')$. Finally, we show that t^* is a regular point of $\hat{x}$. We take a sufficiently small scaling factor a'' with $0 < a'' < a'$ so that the cutoff function

$$(t_1, t_2) \mapsto h((t_1 - t_1^*)/a'', (t_2 - t_2^*)/a'')$$

has support contained in $S_{\alpha'}(t^*)$. The localized image

$$h\big((t_1 - t_1^*)/a'', (t_2 - t_2^*)/a''\big) \cdot \hat{x}(t_1, t_2) \tag{38}$$

is then a product of a uniformly Lipschitz α image with a smooth cut-off function, and is hence a uniformly Lipschitz α function with regularity $\geq \alpha$. Hence, t^* is a regular point of $\hat{x}$, which finishes the proof.

□

B. Experiments

In this section, we give the supplementary material for our experiments in Section 7.

B.1. Implementation Details

We implemented our methods and experiments in PyTorch [32] and describe the details for each method in the following.

ShearletX

Our implementation of the shearlet transform is an adaptation of the python library pyShearLab2D[3] to PyTorch. The digital shearlet coefficients of a $3 \times 256 \times 256$ image are returned as a $49 \times 3 \times 256 \times 256$ tensor where the first 49 channels capture the discretely sampled scale and shearing parameters of the shearlet transform. To optimize the shearlet mask on the $49 \times 3 \times 256 \times 256$ tensor, we use the Adam optimizer [23] with learning rate 10^{-1} and for the other Adam parameters we use the PyTorch default setting. The mask is optimized for 300 steps. The expectation in the ShearletX optimization objective

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

is approximated with a simple Monte Carlo average over 16 samples from ν, which samples uniform noise adapted to each scale and shearing parameter. More precisely, the perturbation for scale a and shearing s is sampled uniformly from $[\mu_{a,s} - \sigma_{a,s}, \mu_{a,s} + \sigma_{a,s}]$ where $\sigma_{a,s}$ and $\mu_{a,s}$ are the empirical standard deviation and mean of the image's shearlet coefficients at scale a and shearing s. The mask is initialized with all ones as in [24]. For the hyperparameters λ_1 and λ_2, we found $\lambda_1 = 1$ and $\lambda_2 = 2$ to work well in practice but many other combinations are possible if one desires more or less sparse explanations.

WaveletX

The discrete wavelet transform (DWT) returns approximation coefficients and detail coefficients. The detail coefficients are parametrized by scale and by orientation (vertical, horizontal, and diagonal). The number of DWT coefficients is the same as the number of pixels and WaveletX optimizes a mask on the DWT coefficients. For the implementation of the DWT, we use the PyTorch Wavelets package[4]. The mask on the DWT coefficients is optimized with the Adam optimizer [23] with learning rate 10^{-1} and for the other Adam parameters we use the PyTorch default setting. The mask is optimized for 300 steps. The expectation in the WaveletX optimization objective

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

approximated with a simple Monte Carlo average over 16 samples from ν, which samples uniform noise adapted to each scale of the wavelet coefficients, analogous to ShearletX. More precisely, the perturbation for scale a is sampled uniformly from $[\mu_a - \sigma_a, \mu_a + \sigma_a]$ where σ_a and μ_a are the empirical standard deviation and mean of the image's wavelet coefficients at scale a. The mask is initialized with all ones as in ShearletX and in [24]. For the hyperparameters, λ_1 and λ_2 we found $\lambda_1 = 1$ and $\lambda_2 = 10$ work well in practice but many other combinations are possible if one desires more or less sparse explanations.

CartoonX

For the examples in the CartoonX method from [24], we used the same parameters and procedure as WaveletX but set $\lambda_2 = 0$. This is because CartoonX and WaveletX only differ in the new spatial penalty that is controlled by λ_2.

Smooth Pixel Mask

For the smooth pixel mask method by Fong et al. [13], we use the TorchRay[5] library, which was written by Fong et al. [13]. The only hyperaparameter for smooth pixel masks is the area constraint, where we use only the values 20%, 10%, or 5%, as did Fong et al. [13].

Pixel Mask without Smoothness Constraints

The pixel mask method without smoothness constraints has the following optimization objective:

$$\max_{m\in[0,1]} \mathop{\mathbb{E}}_{u\sim\nu}\Big[\Phi_c(x \odot m + (1-m)\odot u)\Big] - \lambda \cdot \|m\|_1,$$

The mask m on the pixel coefficients is optimized with the Adam optimizer [23] with learning rate 10^{-1} and for the other Adam parameters we use the PyTorch default setting. The mask is optimized for 300 steps. The expectation in the optimization objective is approximated with a simple Monte Carlo average over 16 samples from ν, which is chosen as uniform noise from $[-\sigma + \mu, \mu + \sigma]$, where μ and σ are the empirical mean an standard deviation of the pixel values of the image, as in [24].

Edge Detector

For the edge detector, we use a shearlet-based edge detector, introduced by Reisenhofer et al. in [34] and adapt the imlementation (PyCoShREM[6] library) by Reisenhofer et al. in [34] to PyTorch. We used the shearlet-based edge detector because it was able to extract edges more reliably than a Canny edge detector [7] and is mathematically well-founded.

[3] https://na.math.uni-goettingen.de/pyshearlab/pyShearLab2D.m.html

[4] https://pytorch-wavelets.readthedocs.io/en/latest/readme.html

[5]]https://github.com/facebookresearch/TorchRay

[6] https://github.com/rgcda/PyCoShREM

B.2. Runtime

In Table 1, we compare the runtime of ShearletX and WaveletX for the ImageNet classifier MobilenetV3Small [21] to (1) smooth pixel mask [13], (2) pixel attribution methods, such as, Guided Backprop [41], Integrated Gradients [42], and Grad-CAM [37], and (3) LIME [35]. All mask explanations, *i.e.*, smooth pixel masks, WaveletX, and ShearletX, are much slower than the pixel attribution methods, which only use a single backward pass for the explanation. ShearletX is roughly 5× slower than smooth pixel masks and WaveletX, because the shearlet mask has more entries than pixels or wavelet coefficients and the shearlet transform involves more computations. In the future, we are eager to significantly speed-up ShearletX by optimizing our implementation, using a less redundant shearlet system to reduce the numbner of coefficients of the mask, and exploring better initialization strategies for the shearlet mask to obtain faster convergence. For instance, we hope to train a neural network in the future that outputs mask initializations for ShearletX that lead to faster convergence.

Method	Time
Integrated Gradients [42]	0.31s
Guided Backprop [41]	0.13s
Grad-CAM [37]	0.13s
LIME [35]	5.22s
Smooth Mask [13]	11.61s
WaveletX (ours)	7.99s
ShearletX (ours)	54.26s

Table 1. Computation time for explanation of MobilenetV3Small [21] decision on ImageNet [11]. It is well-known that mask explanations are more computationally expensive than pixel attribution methods, such as Integrated Gradients [42], Grad-CAM [37], and Guided Backprop [41]. ShearletX is slower than WaveletX and Smooth Pixel Masks [13] due to the mask on the shearlet representation being larger and shearlets involving more computations.

B.3. Scatter Plots

The scatter plots in Figure 4 in the main paper compares the hallucination score and conciseness-preciseness score between ShearletX, WaveletX, smooth pixel masks by Fong et al. [13], and pixel masks without smoothness constraints. In this section, we provide evidence that our results from Figure 4 are consistent across different area constraints for smooth pixel masks and across different classifiers. In Figure 8, we show the scatter plots for Resnet18 [17] for the area constrains 5%, 10%, and 20%. Figure 9 shows the same plots for a MobilenetV3Small [21] network.

B.4. Quantitative Comparison

Pixel attribution methods, such as Integrated Gradients [42], Guided Backprop [41], and Grad-CAM [37], are commonly compared by insertion and deletion curves [4, 24, 33, 36], which gradually insert/delete the most relevant pixels and observe the change in class probability. A good insertion curve exhibits a rapid initial increase in class probability and large area under the curve. A good deletion curve exhibits a rapid initial decay and small area under the curve. Comparing ShearletX on insertion and deletion curves poses two challenges: (1) ShearletX is given by a mask that is defined in shearlet space and not in pixel space, as in other methods. (2) ShearletX does not give a proper ordering for the relevance of coefficients due to the binary nature of the mask.

In Figure 6, we compare insertion and deletion curves for ShearletX, where we perturb the most relevant coefficients for ShearletX in *shearlet space*. Insertion and deletion curves are averaged over 50 random ImageNet validation samples and compared on MobilenetV3Small [21], ResNet18 [17], and VGG16 [39]. ShearletX performs best among compared methods on the *initial part* of the insertion curves in Figure 6, exhibiting a rapid initial increase in probability score. This is what ShearletX was optimized for: keeping very few coefficients that retain the classification decision. However, once ShearletX achieves its peak, coefficients are inserted that were probably marked as zero and not further ordered by the shearlet mask and therefore inserted in arbitrary order. Consequently, the probability score collapses after the peak. Similar behavior is observed for the deletion curve, which initially decays rapidly for ShearletX and then slows down due to the lacking ordering of unselected coefficients. The hyperparameters for ShearletX in the experiments of Figure 6 are ($\lambda_1 = 1$ and $\lambda_2 = 2$).

In Figure 7, we also experiment with a pixel ordering for ShearletX by ordering pixels simply by their magnitude in the ShearletX explanation. Surprisingly, this ordering beats all other compared methods on the insertion curves on two out of three classifiers that we evaluate. The deletion curves for the pixel ordering of ShearletX are competitive but not outperforming the other methods. For the pixel ordering of ShearletX, we used smaller sparsity parameters ($\lambda_1 = 0.5$ and $\lambda_2 = 0.5$) to avoid having too many deleted pixels in ShearletX that cannot be ordered uniquely.

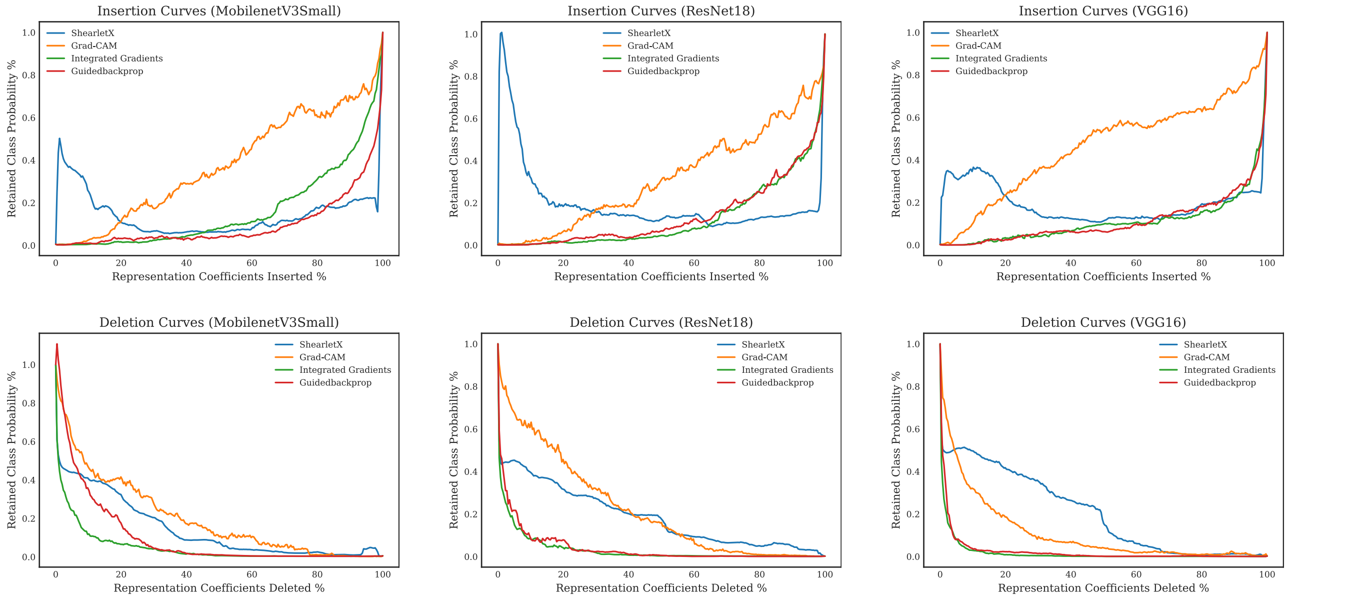


Figure 6. Insertion and deletion curves for ShearletX and popular pixel attribution methods (Integrated Gradients [42], Grad-CAM [37], and Guided Backprop [41]), where representation coefficients are flipped and set to zero. For ShearletX, the representation coefficients are the shearlet coefficients and for all other methods, the pixel coefficients. Insertion curves plot the percentage of inserted representation coefficients (the most relevant coefficients first) against the retained class probability (class probability after perturbing divided by original class probability). Deletion curves plot the percentage of deleted representation coefficients (the most relevant coefficients first) against the retained class probability. First row: Insertion curves for MobilenetV3SMall [21], ResNet18 [17], and VGG16 [39]. Second row: Deletion curves for MobilenetV3SMall [21], ResNet18 [17], and VGG16 [39].

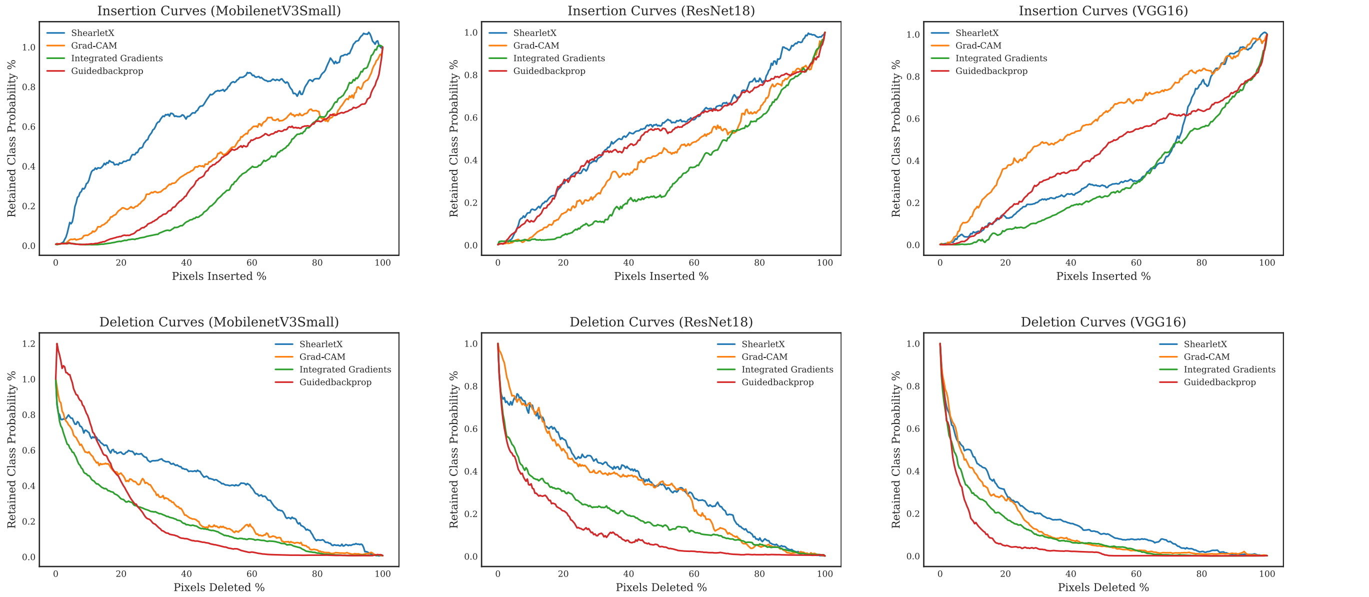
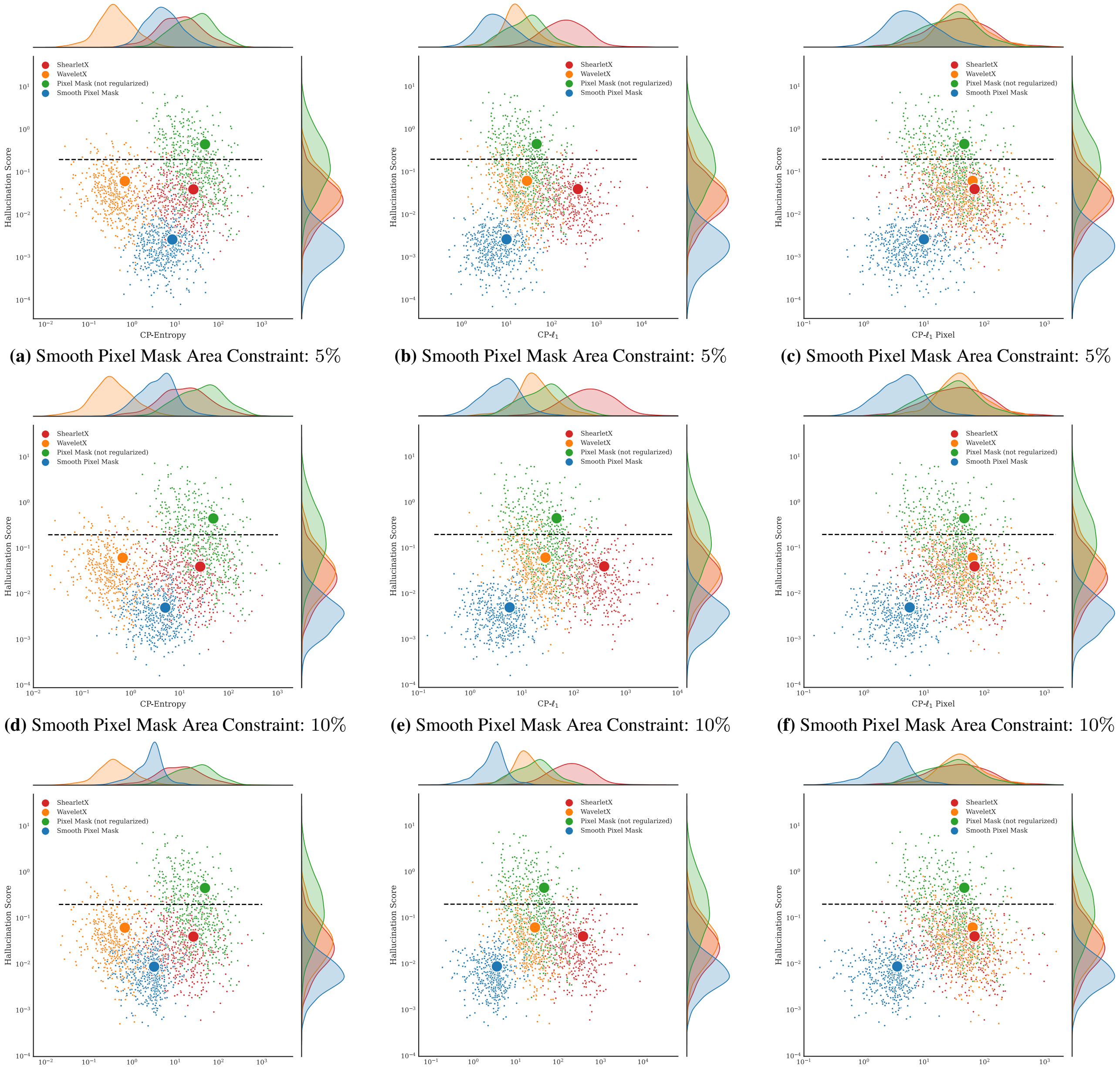


Figure 7. Insertion and deletion curves for ShearletX and popular pixel attribution methods (Integrated Gradients [42], Grad-CAM [37], and Guided Backprop [41]). For ShearletX, we sort pixels by magnitude in the explanation. Insertion curves plot the percentage of inserted pixels (the most relevant pixels first) against the retained class probability (class probability after perturbing divided by original class probability). Deletion curves plot the percentage of deleted pixels (the most relevant pixels first) against the retained class probability. Deleted pixels are replaced with blurred pixel values. First row: Insertion curves for MobilenetV3SMall [21], ResNet18 [17], and VGG16 [39]. Second row: Deletion curves for MobilenetV3SMall [21], ResNet18 [17], and VGG16 [39].



(a) Smooth Pixel Mask Area Constraint: 5%

(b) Smooth Pixel Mask Area Constraint: 5%

(c) Smooth Pixel Mask Area Constraint: 5%

(d) Smooth Pixel Mask Area Constraint: 10%

(e) Smooth Pixel Mask Area Constraint: 10%

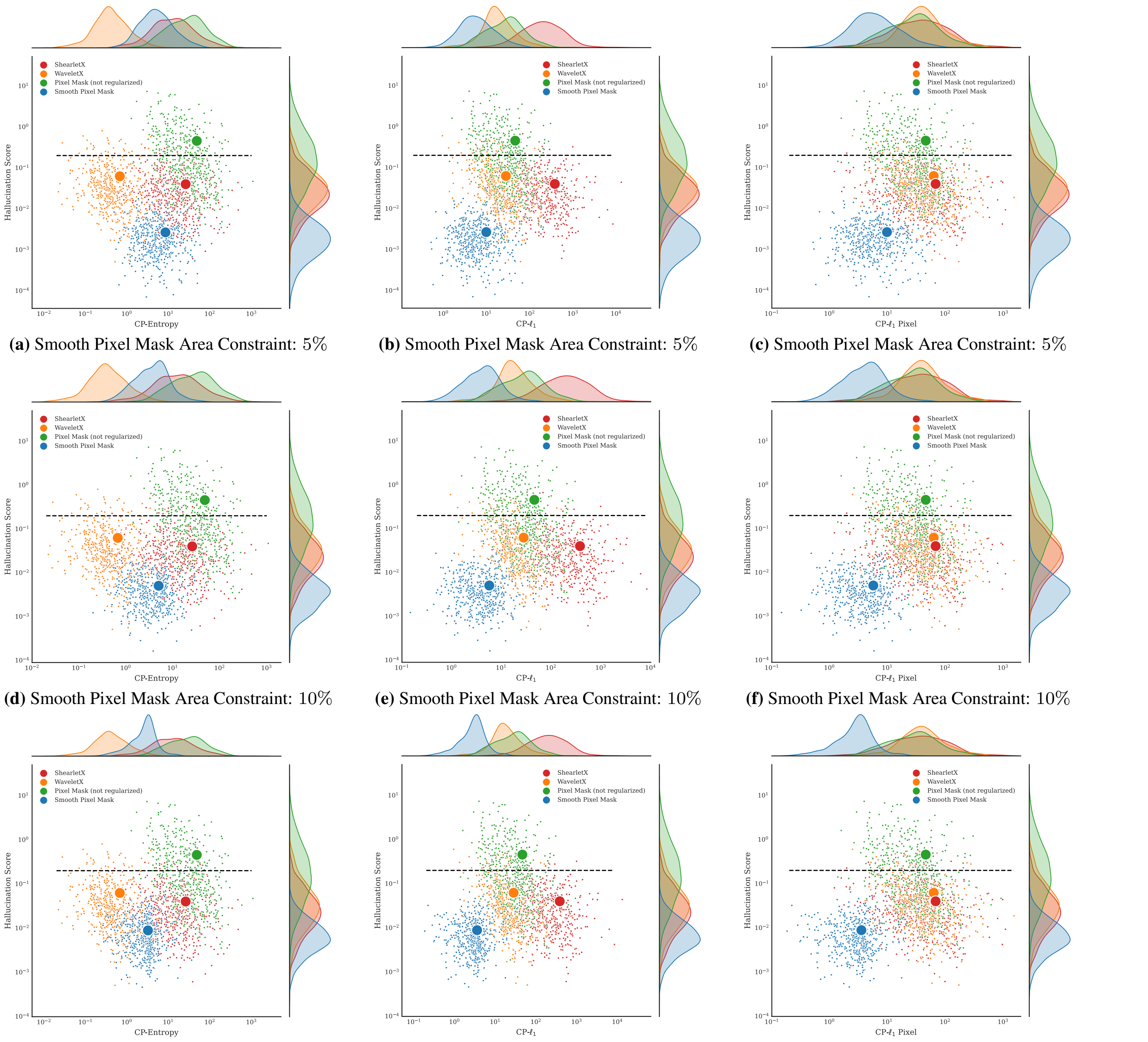
(f) Smooth Pixel Mask Area Constraint: 10%

(g) Smooth Pixel Mask Area Constraint: 20%

(h) Smooth Pixel Mask Area Constraint: 20%

(i) Smooth Pixel Mask Area Constraint: 20%

Figure 8. Scatter plots 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. We used the classifier ResNet18 [17] for all scatter plots. First row uses smooth pixel masks [13] with area constraint 5%, second row uses 10%, and last row uses 20%. The scatter plots shows that the advantage of ShearletX over smooth pixel masks [13] holds for different area constraints.



(a) Smooth Pixel Mask Area Constraint: 5%

(b) Smooth Pixel Mask Area Constraint: 5%

(c) Smooth Pixel Mask Area Constraint: 5%

(d) Smooth Pixel Mask Area Constraint: 10%

(e) Smooth Pixel Mask Area Constraint: 10%

(f) Smooth Pixel Mask Area Constraint: 10%

(g) Smooth Pixel Mask Area Constraint: 20%

(h) Smooth Pixel Mask Area Constraint: 20%

(i) Smooth Pixel Mask Area Constraint: 20%

Figure 9. 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. We used MobilenetV3Small [21] as a classifier for all scatter plots. First row uses smooth pixel masks [13] with area constraint 5%, second row uses 10%, and last row uses 20%. The scatter plots shows that the advantage of ShearletX over smooth pixel masks [13] holds for different area constraints. The scatter plots compared to Figure 8 also show that the advantage of ShearletX over smooth pixel masks [13] holds for different classifiers.