Inpainting Insights: Elevating Visual XAI with Photorealistic Perturbations

Josef Lindl[1][0009−0007−6656−0632], Mariana Chaves[2][0009−0001−8086−9034], and Damien Garreau[1][0000−0002−7855−2847]

[1] Julius-Maximilians-Universität Würzburg, Sanderring 2, 97070 Würzburg, Germany
lindl.josef@googlemail.de
damien.garreau@uni-wuerzburg.de
[2] University of Groningen: Broerstraat 5, 9712 CP Groningen, The Netherlands
m.e.chaves.espinoza@rug.nl

Abstract. The increasing complexity of state-of-the-art machine learning models has made their behavior progressively harder to interpret, spurring rapid advancements in the field of eXplainable Artificial Intelligence (XAI). Among many methods proposed, perturbation-based approaches play a major role. By systematically altering (perturbing) input features, these approaches measure the impact on the model's predictions. For image data, traditional perturbation techniques, often involve replacing pixel values *e.g.*, with a pre-defined color. However, such approaches, but also more refined deterministic techniques, generate unrealistic out-of-distribution samples and often leave visible artifacts, which can mislead the model and compromise explanation quality. In this work, we adjust LIME, a widely used perturbation-based method, to demonstrate how generative inpainting can improve perturbation-based explanations for images. We achieve photorealistic perturbed samples that align better with the original data distribution and enhance explanation quality.

Keywords: Perturbation-based Explanations · Generative Inpainting.

1 Introduction

As Artificial Intelligence (AI) systems increasingly influence critical decision-making, the need for transparency and accountability has become paramount [20, 22, 5]. For example computer vision models can achieve high accuracy by exploiting unintended dataset artifacts rather than learning meaningful visual features [16]. The research field of eXplainable AI (XAI) explores methods, to make complex machine learning models understandable to humans, fostering trust and ethical deployment.

Perturbation-based XAI methods, such as LIME (Local Interpretable Model-agnostic Explanations) [23], SHAP (SHapley Additive exPlanations) [17], and RISE (Randomized Input Sampling for Explanation) [21], have become foundational tools for interpreting image classifiers [4]. These approaches generate

perturbed samples by systematically removing or occluding parts of the input image, replacing them with a constant color (*e.g.*, gray, or mean pixel value), blurring, or zeroing out pixels [23, 34, 21].

As an example, LIME for images segments the input image into superpixels and creates perturbed samples through occlusion, replacing them with the mean color of the superpixel by default. However, LIME perturbation sampling can be adopted for different replacement methods, as we demonstrate in Section 2.2 by 8 deterministic occlusion strategies. Still, we can show, that there are cases where all of these approaches fail. More generally, despite being widely used, deterministic perturbation methods tend to introduce unnatural artifacts and generate out-of-distribution samples. This can lead to unpredictable model behavior and ambiguous attributions [2, 29].

A shared intuition behind replacement approaches is that occlusion should create plausible content in place of the removed region. Agarwal and Nguyen [1] demonstrated that generative inpainting models can maintain the original data distribution while removing discriminative features. Their adaption of LIME, LIME-G, uses inpainting with DeepFill v1 and v2 [33, 32] to generate perturbed samples. However, more recent models, such as LaMa [25], outperform DeepFill in perceived inpainting quality. Nevertheless, despite these advancements, no work has yet explored improving LIME-G with state-of-the-art inpainting models.

Our contributions are as follows: Firstly, we introduce **Leveraging Inpainting for Local Interpretability (LILI)**, a method that incorporates generative perturbation mask inpainting to explain images using LIME. Secondly, by leveraging overlapping superpixels (perturbation mask expansion), we obscure residual artifacts and outlines for more effective occlusion. We demonstrate that mask expansion helps mitigate the unintended reconstruction of occluded features by adapting the perturbation process to the generative capabilities of modern inpainting models. Finally, our evaluation reveals that LILI produces more realistic perturbations regarding perceived image quality, favoring more consistent distributions of high-level features deep within the explained network. Moreover, we show that LILI achieves improved explanation quality compared to LIME and LIME-G, highlighting the effectiveness of using advanced generative inpainting models for perturbation-based explanations.

Fig. 1 shows an example of perturbed samples, as well as explanations as given by LIME-G, and LILI. Note that only LILI identifies the band-aid as most important. All the code for the method and the evaluation is open source and available on GitHub[3,4].

[3] https://github.com/jo01123/LILI
[4] https://github.com/m-chaves/LIME

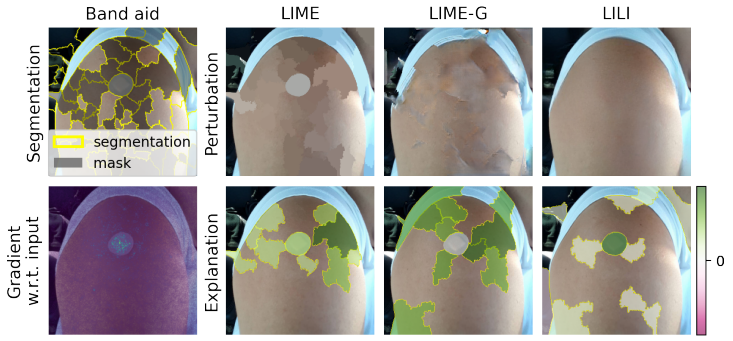

Fig. 1. Perturbed samples and explanations for *band-aid* from different methods: The first row shows the segmented *original* image with an example of a perturbation mask (left image), as well as the perturbed samples as given by *LIME*, *LIME-G* and *LILI*. The second row shows the model's gradients with respect to the input, as well as the explanations given by different methods (here, the top ten superpixels by absolute weight are visualized: darker green indicates a stronger positive contribution). Note that the gradient w.r.t. the input reveals that the band-aid superpixel seems to influence the prediction. However, only LILI identifies the band-aid as most important.

2 Background and Motivation

In this section, we will outline the foundations for the proposed method LILI. Firstly, we introduce LIME in Section 2.1. Secondly, in Section 2.2, we discuss examples, where deterministic occlusion fails. Section 2.3 introduces the LaMa inpainting model and discusses why it was selected for LILI. Finally, in Section 2.4, we present related work.

2.1 LIME

Ribeiro et al. [23] introduced Local Interpretable Model-agnostic Explanations (LIME), a method for explaining individual predictions (local) of any black-box model (model-agnostic). The core idea is to approximate the complex model locally with an interpretable surrogate model, such as linear regression. For this, LIME perturbs the input data by modifying interpretable features and observes the resulting changes in the model's output. The surrogate model is then fitted to these changes to estimate feature importance. For image data, LIME defines interpretable features as superpixels, groups of pixels with similar properties. Perturbed samples are created by randomly masking out some of these superpixels, allowing the surrogate model to approximate the behavior of the original model around the specific instance being explained.

Following the notation of Garreau and Mardaoui [10], we can describe the core steps of LIME for images as follows. Let f be the model and ξ the sample to explain.

1. *Create interpretable features* by segmenting the image into d disjunct superpixels $J_1, \ldots, J_d$ using a segmentation algorithm (*e.g.*, quickshift [30] by default).
2. *Create perturbed samples* $x_1, \ldots, x_n$. For each x_i draw $z_i \in \{0,1\}^d$ from independent Bernoulli trials. For all $z_{i,j} = 1$ the superpixel J_j is masked for sample x_i *i.e.*, switched off. Replace therefore all pixel of masked superpixels with the superpixel's mean color (default) or a replacement color. An example of a segmentation, a perturbation mask and a perturbed sample is given in Fig. 1.
3. *Compute weights* w_i *from* x_i*s proximity to* ξ *e.g.*, using an exponential kernel. Perturbed samples closer to the original (*i.e.*, with less superpixels switched off) receive a higher weight.
4. *Query the model* by making predictions $y_i = f(x_i)$.
5. *Fit the local surrogate model* to the y_is (target) on the presence or absence of superpixels (z_is). By default, Ridge regression weighted by w_i is used. The resulting coefficients of the surrogate model can be interpreted as the approximated contributions of the corresponding superpixels to the prediction of the model f on the original image ξ. By mapping the pixels of the image to its superpixel weights, we receive the attribution map, which can be visualized as heatmap.

2.2 Where Deterministic Occlusion Fails

A natural idea is to improve superpixel replacement with a better, but simple method. For instance, we tested eight deterministic approaches for LIME (including the native LIME replacement by mean superpixel color or black): (1) mean color of the superpixel, (2) mean image color, (3) mean color of neighboring superpixels, (4) zeroing out (black) (5) median color of pixels at the image border (6) median color of superpixels at the image border (7) median color of superpixel's neighbors mean, and (8) median color of the superpixel means.

While these methods showed improvements in specific cases, they still produce visually unnatural samples and fail to generalize effectively. LIME struggles to identify the feature's importance, when the replacement color closely resembles the original superpixel, or if the shape of the superpixel reveals information about its content. We present a result by the example of the *band-aid* class: In Fig. 2 none of these strategies identifies the band-aid as most important superpixel. LILI, however, does (see Fig. 1). Additional results can be found on GitHub[4].

2.3 Resolution-robust Large Mask Inpainting with Fourier Convolutions (LaMa)

LaMa, introduced by Suvorov et al. [25], is a feed-forward inpainting network designed to handle high-resolution images, large missing areas, and complex

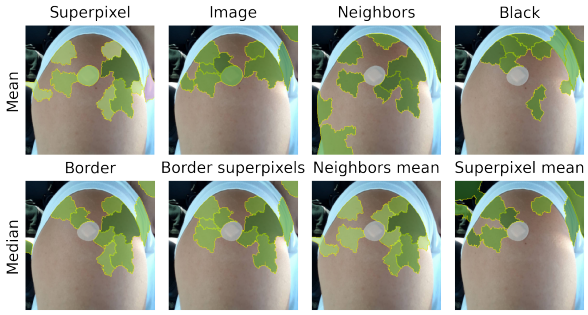

Fig. 2. Explanations using 8 deterministic occlusion methods: Applied to an image of class *band-aid* where default LIME fails to identify the most important superpixel(s). The top ten superpixels by absolute weight are visualized (darker green for a stronger positive contribution). From left to right, we can see the occlusion methods 1-8 as described in Section 2.2.

geometric shapes by using fast Fourier convolutions (FFC) [6]. FFC split the input channel-wise to a local and a global branch and use channel-wise fast Fourier transformations on the global branch for an image width receptive field. The model architecture of LaMa is similar to a ResNet with 3 downsampling blocks, followed by 6-18 FFC residual blocks, followed by 3 upsampling blocks. LaMa is trained on the Places and the CelebA-HQ datasets using real images and synthetically generated masks, and loss function, that is a weighted sum of a gradient penalty, an adversarial loss, a discriminator-based perceptual loss, and a newly introduced loss which they call high receptive field perceptual loss. Despite its relatively lightweight architecture (27 million parameters), LaMa consistently outperforms older models like DeepFill v2 [33] in perceptual quality metrics such as Fréchet Inception Distance (FID) [11] and LPIPS [35].

We chose LaMa for inpainting due to its balance between perceptual quality and computational efficiency, making it well-suited for generating the large number of perturbed samples required by LIME (typically 1000 per explanation). While state-of-the-art inpainting models based on transformers [8] or diffusion architectures [13] achieve superior perceptual quality, they are computationally expensive. For instance, even the region-aware diffusion model (RAD) [13], which is faster than earlier diffusion-based methods, has significantly higher inference times compared to LaMa.

2.4 Related Work

Perturbation-based explanation methods are widely used in XAI for computer vision [4]. A key challenge in these methods is generating perturbed samples that remain realistic and align with the original data distribution [2, 29]. Recent ad-

vancements in generative inpainting have shown promising results in addressing this issue by creating realistic perturbations that preserve contextual consistency.

Tritscher et al. [29] apply generative inpainting to Shapley-value-based explanations for anomaly detection with tabular data. By replacing traditional feature occlusion methods with a denoising diffusion probabilistic model (TabDDPM [15]), they demonstrate stable explanation quality across various datasets and models.

Regarding image data, Agarwal and Nguyen [1] and Aghakishiyeva et al. [2] explore the capabilities of generative inpainting for the generation of perturbed samples. By introducing LIME-G, Agarwal and Nguyen [1] integrate generative inpainting into LIME. LIME-G replaces per superpixel mean-color occlusion with DeepFill-v1/v2 [33, 32], to generate perturbed samples that are plausible and better aligned with the original data distribution. Their findings showed that inpainting-based perturbations improve explanation accuracy and robustness compared to traditional occlusion methods. Likewise, Aghakishiyeva et al. [2] explore photorealistic inpainting for perturbation-based explanations in ecological monitoring. Their method uses refined masks from the Segment Anything Model (SAM) [14] and Stable Diffusion [24] to replace or remove objects or background in drone images. However, the computational cost of this method using SAM and Stable Diffusion, limits it's scalability.

Building on the work of Agarwal and Nguyen [1], LILI leverages LaMa (a more recent inpainting model) and mask expansion to generate more realistic perturbed samples with enhanced occlusion. Compared to Aghakishiyeva et al. [2] LILI balances computational cost and realism, as LaMa is less computationally expensive than Stable Diffusion.

3 Leveraging Inpainting for Local Interpretability

As discussed previously, LILI adopts LIME as given by the original implementation (see Section 2.1) and only adjusts the second step of the explanation process: the creation of perturbed samples. Two major changes are applied for LILI. Firstly, the generative replacement of superpixels by the LaMa inpainting model, which is explained in Section 3.1. Secondly, the introduction of a hyperparameter m to enable the expansion of perturbation masks, which is discussed in Section 3.2.

3.1 Perturbation by inpainting

Regarding the generative replacement of superpixels for perturbed samples for LIME, we proceed as follows. Let the image be of height H and width W. Further let $M^{(i)} \in \{0,1\}^{H \times W}$ be the perturbation mask with $M^{(i)}_{k,\ell} = 1$ iff the superpixel corresponding to index (k, ℓ) is masked (*i.e.*, switched of). LIME, if superpixel J_j is switched off, replaces all pixels of J_j with the superpixel's mean color (default). LIME-G and LILI however, generate the perturbed samples x_i

using the inpainting model $\mathcal{I}$, the original image ξ and the mask $M^{(i)}$:

$$x_i = \mathcal{I}(\xi, M^{(i)}). \tag{1}$$

For LIME-G $\mathcal{I}$ is given by DeepFill, for LILI by LaMa respectively. In our experiments LIME-G is reimplementated using the PyTorch implementation of the DeepFill model [3].

3.2 Perturbation mask expansion

In this section, we will first motivate the establishment of the hyperparameter m, before we explain the technical details of the method realization.

The introduction of the hyperparameter m for mask expansion addresses limitations arising from the inpainting of superpixels generated by LIME's default segmentation algorithm. Segmentation methods, such as quickshift, tend to leave small artifacts at the border of objects [9], which can provide unintended cues to the inpainting model (see Fig. 3). For LaMa these artifacts can be sufficient to reconstruct the occluded object with high perceptual accuracy, contradicting the intended removal of features for perturbation. By expanding the perturbation mask by m pixels around the original superpixel boundaries, these artifacts can be hidden, forcing the inpainting model to generate content that aligns more closely with the background. Fig. 3 illustrates this effect: The masked area here includes three *rose hips*. For effective feature occlusion we would expect them to be replaced (*e.g.*, with a plausible background). Without mask expansion, however, LaMa reconstructs the rose hips based on residual artifacts. By using mask expansion, on the other hand, the rose hips are successfully removed. The right part of the figure shows more examples of occlusion using the different methods. Looking at the classification confidence (annotated below) reveals that the occlusion using LILI with perturbation mask expansion, results in a significant drop. Supported by these examples, we hypothesize that this approach enhances the effectiveness of occlusion for perturbation-based sample generation.

Regarding the hyperparameter for perturbation masks expansion, we proceed as follows. Instead of computing the mask expansion for each perturbation mask, we compute the expanded masks $E^{(J_j)} \in \{0,1\}^{H \times W}$ for every superpixel J_j. This can be done prior to the sampling process and is computationally more efficient - as usually the number of perturbed samples is much larger than the number of superpixels. The algorithm follows three major steps: With m pixel mask expansion, we compute the expanded superpixel masks as follows. For $J_j \in J_1, \ldots, J_d$:

1. Find the contour C of J_j using the border following algorithm of Suzuki and Abe [26].
2. Set $E^{(J_j)}_{k,\ell} = 1$ iff index $(k,\ell) \in J_j$.
3. Follow the contour for $(k,\ell) \in C$ and set $E^{(J_j)}_{r,s} = 1$ for every pixel index (r,s) in the square of m pixel around (k,ℓ).

The final expanded perturbation mask $\mathcal{M}^{(i)}_{k,\ell} \in \{0,1\}^{H \times W}$ for the ith perturbed sample is given by the element-wise disjunction (logical or) of the ex-

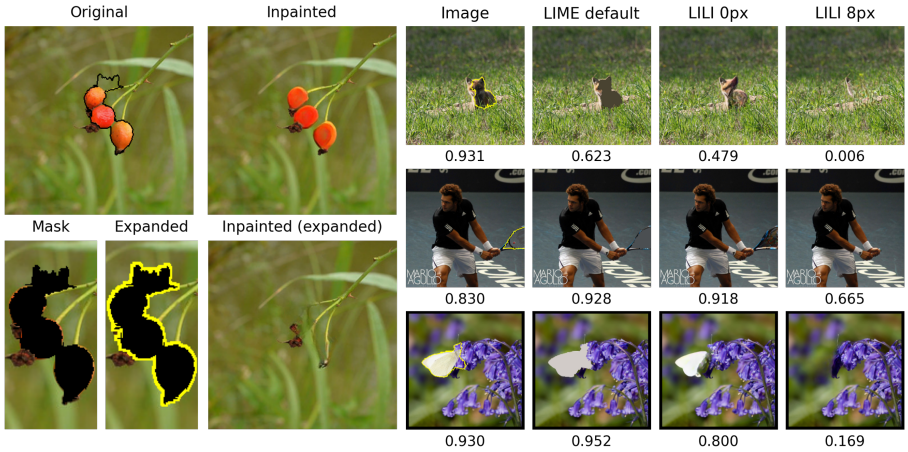

Fig. 3. Motivation for perturbation mask expansion: The top-left image (*original*) shows three hips and a perturbation mask (black outline). The *inpainted* sample closely resembles the original, but the *mask* does not fully cover the hips. The *expanded* mask hides remaining artifacts, encouraging background replacement (see *inpainted expanded*). The right part shows perturbation effects on classification confidence (probabilities below). The *image* column displays the original and masked region (*i.e.*, yellow outline). *LIME default* and *LILI* with *0 pixel* expansion leave the instance visible with high confidence, while *LILI* with *8 pixel* expansion removes it, causing a significant confidence drop.

panded superpixel masks $E^{(J_j)}$ of all superpixels J_j that are switched off (where $z_{i,j} = 1$).

4 Evaluation

The proposed method, LILI, is evaluated alongside LIME and LIME-G across several criteria: inpainting quality (Section 4.1), explanation quality (Section 4.2), explanation stability (Section 4.3), and computation time (Section 4.4). For the generation of explantions we generally use the default LIME hyperparameters, with adjustments for generating perturbed samples for LILI and LIME-G as described in Section 3. Images are selected randomly from the ImageNet 1k dataset of the ILSVRC 2012 challenge [27]. For each image, we explain the class with the highest probability as predicted by the InceptionV3 model [27], trained on the ImageNet-1k dataset.

4.1 Inpainting Quality

The LaMa model, used in LILI, consistently outperforms DeepFill (used in LIME-G) in perceptual inpainting quality across datasets and metrics [25]. To

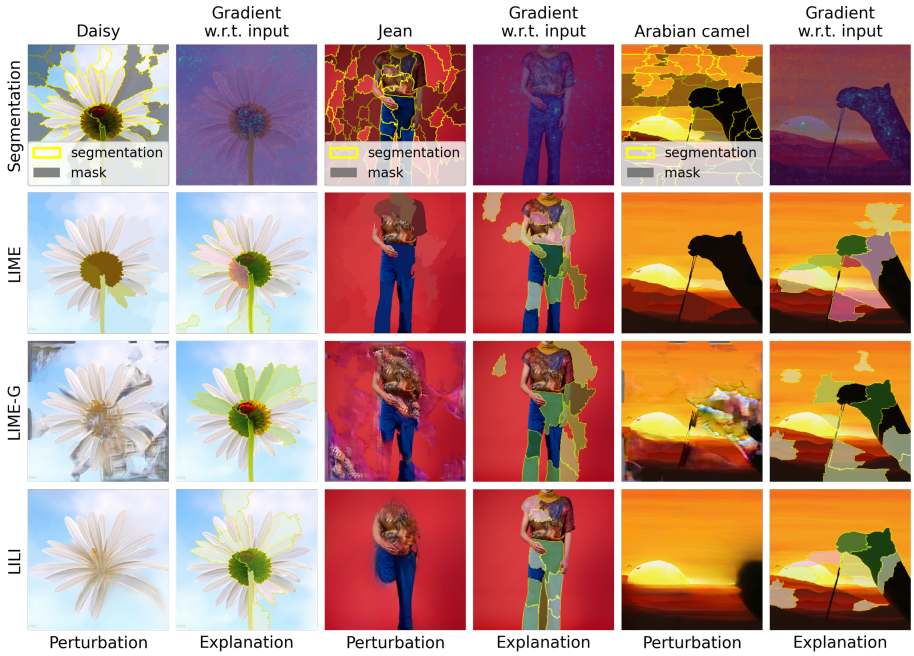

Fig. 4. Examples of perturbed samples and explantions by methods: For each explained image, we can see the original with it's *segmentation* and a perturbation mask on the left, and the gradients with respect to the input on the right (first row). Following rows show one perturbed sample, and the explanation as given by the corresponding methods.

validate this for LIME perturbation mask inpainting, we conducted evaluations using the Fréchet Inception Distance (FID) [11]. The FID compares the distributions of high-level image features given by the activations of a classification CNN to quantify the dissimilarity between real images and inpainted perturbed samples. It has been shown empirically, that the FID correlates with human perception of image quality [11].

As originally proposed [11], we compute the activations as given by the 2048-dimensional activation vectors of the last pooling layer of the InceptionV3 network using the PyTorch FID package [19]. The FID (see Eq. 2) is then given by the Fréchet distance between the multidimensional Gaussian distributions fitted to the activations from real images $\mathcal{N}(\mu, \Sigma)$ and generated images $\mathcal{N}(\mu_g, \Sigma_g)$, where μ and Σ are the estimated mean vector and covariance matrix of the distribution of real images (μ_g, Σ_g of the generated distribution respectively):

$$d_F\left(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu_g, \Sigma_g)\right)^2 = \|\mu - \mu_g\|_2^2 + \operatorname{trace}\left(\Sigma + \Sigma_g - 2\left(\Sigma\Sigma_g\right)^{\frac{1}{2}}\right). \quad (2)$$

Note that we intentionally used the same model (InceptionV3) to compute both the FID and the explanations. While Agarwal and Nguyen [1] advocate higher dissimilarity for occlusion (MS-SSIM [31] and LPIPS [35]), we favour a lower FID, which indicates superior perceived quality and tighter internal activation alignment between real images and perturbed samples, thereby minimizing the risk of prediction-biasing artifacts.

We compute the FID from a reference set of 10,000 real images and compare it to a second set, that consists of 5 perturbed samples for each of the real images (50,000 images total). We repeat this for LILI, LIME-G and LIME perturbations. Table 1 shows that LILI achieves the lowest FID score (6.727), significantly outperforming LIME-G (56.524) and LIME (30.532).

Table 1. Fréchet inception distance (FID) of perturbed samples: As generated by LIME, LILI and LIME-G. Lower values indicate a higher perceptual image quality.

	Occlusion method	FID
LIME	superpixel mean color	30.532
LIME-G	inpainting with DeepFill	56.524
LILI	inpainting with LaMa	<u>6.727</u>

4.2 Explanation Quality

As a proxy for explanation quality, we use the saliency metric [7]. This metric measures how well attribution maps identify regions critical to the model's predictions by computing the smallest rectangular patch containing the most salient regions and measuring the classification probability for the cropped patch. A detailed description follows:

For thresholds $t = \alpha\mu_{max}$ given by $\alpha \in \{0,\ 0.05,\ 0.1,\ 0.15,\ \ldots,\ 0.95\}$ and the maximum of the superpixel weights μ_{max}, the smallest rectangular image patch is cropped, such that no saliency above the threshold is outside the patch. Now the ratio a_t of the area of the patch over the area of the image is computed. Afterwards, the patch is resized to the models input size (independently of the previous aspect ratio) and the classification probability $f(x_t)$ of the upsampled patch x_t is computed. Finally, the saliency metric is given as:

$$s(a_t, f(x_t)) = \log(\max(a_t, 0.05)) - \log(f(x_t)). \tag{3}$$

Interpreted from an information theory perspective, this metric can be understood as the concentration of information in the cropped image [7]. Lower saliency scores indicate better explanations.

The experiments were conducted with 100 randomly selected images from the ImageNet-1k test set, using mask expansions of 0, 3, 5, and 8 pixels for both LILI and LIME-G. For every method ν (LIME, LIME-G, LILI), mask expansion

by $m \in \{0, 3, 5, 8\}$, image i and threshold α, the experiment yields a saliency score $S_{\nu,m,i,\alpha}$. We aggregate across thresholds to compute a score $S^{\star}_{\nu,m,i}$:

$$S^{\star}_{\nu,m,i} = \min_{\alpha} S_{\nu,m,i,\alpha}. \tag{4}$$

Table 2 shows that LILI with a 3-pixel mask expansion achieves the best saliency scores across most statistical measures, outperforming both LIME and LIME-G. This demonstrates that LILI provides faithful and robust explanations. Agarwal and Nguyen [1] showed that LIME-G performs on-par with LIME regarding the saliency metric. Our improved results therefore are of particular interest.

Table 2. Statistics of the saliency metric scores: For each method and mask expansion is presented: mean, first quartile (Q1), median, third quartile (Q3), interquartile range (IQR) and median absolute deviation (MAD). Lower values indicate better explanations. The best result (minimum aggregated saliency score) for mean, Q1, median and Q3 is underlined.

		mean	Q1	median	Q3	IQR	MAD
Method	Mask expansion						
LILI	0	-1.766	-2.410	-1.981	-1.159	1.251	0.574
	3	<u>-1.816</u>	-2.427	<u>-2.011</u>	<u>-1.357</u>	<u>1.070</u>	<u>0.544</u>
	5	-1.733	<u>-2.451</u>	-1.917	-1.245	1.206	0.619
	8	-1.692	-2.427	-1.875	-1.189	1.238	0.632
LIME-G	0	-1.637	-2.387	-1.774	-1.013	1.374	0.691
	3	-1.699	-2.427	-1.864	-1.098	1.329	0.701
	5	-1.647	-2.349	-1.737	-1.059	1.290	0.636
	8	-1.744	-2.367	-1.944	-1.187	1.180	0.630
LIME	0	-1.625	-2.308	-1.756	-1.188	1.119	0.559

4.3 Explanation Stability

The stability of explanations was assessed by computing 10 explanations for each of 100 images, using the same superpixel segmentation but varying perturbed samples. Note that for comparability for both LIME-G and LILI, the perturbation mask is expanded by 5 pixels. Similarly to Tan et al. [28], we compute the average Jaccard index $\mathrm{JI}(A, B) = |A \cap B|/|A \cup B|$ between all pairs A, B of sets of the top k superpixel by absolute weight. Further, we report the coefficient of concordance (Kendall's W [12]) for the top k superpixel (by absolute mean weight) as it captures the agreement of rankings across different runs and naturally supports multiple raters. Kendall's W and Jaccard index range from 0 (no agreement/similarity) to 1 (complete agreement/similarity). Figure 5 shows the results computed for the top 1-20 superpixels. Generally, LILI falls between LIME-G and LIME with the highest stability. However, for Kendall's W and larger k, LILI achieves significantly higher explanation stability than LIME-G and can be compared to LIME.

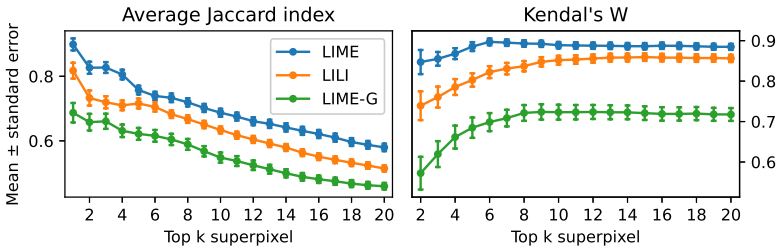

Fig. 5. Average Jaccard index and Kendall's W: Computed for the top 1-20 superpixels by absolute weight for the average Jaccard index and the top 2-20 superpixels by absolute mean weight for Kendal's W. The mean and standard error of the aggregation across all 100 images are visualized.

4.4 Computation Time

The execution times for generating explanations were tracked for 100 images using a NVIDIA L40 GPU for model inference. Table 3 shows that LIME is the fastest method, followed by LIME-G and LILI. The increased runtimes for LILI and LIME-G are attributed to the computational cost of inpainting.

Table 3. Execution times of explanations: The table shows the mean and the standard deviation (std) of the execution times of the generation of explanations.

Method	LIME	LIME-G				LILI			
Mask expansion	0	0	3	5	8	0	3	5	8
mean	4.431	8.708	8.915	9.008	9.154	12.126	12.334	12.425	12.575
std	0.099	0.037	0.044	0.046	0.055	0.214	0.092	0.062	0.073

5 Discussion

While the current experimental setup utilizing LIME's default hyperparameters, the InceptionV3 model, and the saliency metric for ImageNet provides a solid and reproducible baseline, it also opens promising avenues for further exploration. This could be extended by analyzing alternative models, diverse datasets, and more nuanced evaluation metrics to capture a broader spectrum of explanation quality. Since LIME is known to be sensitive to hyperparameter changes [18], future work will in particular include the evaluation of the generalizability of the results for different hyperparameter combinations.

Our results show that LIME-G and LILI produce less stable explanations than LIME, which we hypothesize stems from two main factors: Firstly, for LIME, one masked superpixel is always replaced by one deterministic color

(*e.g.*, superpixel mean). LILI, however, inpaints superpixels depending on the perturbation mask, which is determined by the combination of masked superpixels. While this can be an advantage, if, *e.g.*, an object consists of two separate superpixels being only erased if both are inpainted, it might influence the explanation stability. Secondly, mask expansion introduces overlapping superpixels, which can influence neighboring regions, causing explanations to be less coherent. This highlights the need for careful tuning of the mask expansion parameter to balance effective occlusion and unintended interactions.

Finally, it should be noted that the perturbation mask expansion is roughly equivalent to morphological dilation, an operation that could have been implemented using existing libraries for better performance. However, this optimization is negligible in our setting, as the operation runtime is small compared to the sample generation.

6 Conclusion

In this work, we introduced Leveraging Inpainting for Local Interpretability (LILI), a novel approach to improve perturbation-based explanations by generating more realistic perturbed samples using the LaMa inpainting model. The Fréchet Inception Distance (FID) results confirm that LILI, compared to LIME and LIME-G, produces perturbed samples with higher perceived image quality, that are better aligned with the original data distribution, fostering trust by reducing the risk for unexpected model behaviour. LILI achieves, compared to LIME and LIME-G, superior explanation quality, as demonstrated by the saliency metric, particularly when combined with the proposed hyperparameter for mask expansion. However, while LILI offers improvements in explanation quality, it does so with computational costs that are higher than LIME but remain lower than *e.g.*, diffusion-based approaches, such as Aghakishiyeva et al. [2]. This balance between quality and efficiency makes LILI an alternative for perturbation-based explanations in scenarios where computational resources are limited.

Future work will include additional evaluation metrics and models to rigorously validate explanation quality. We will explore the impact of hyperparameters and integrate state-of-the-art inpainting models. Moreover, extending LILI to other modalities and perturbation-based explanation frameworks will provide further insights into its generalizability and applicability across diverse machine learning tasks.

Acknowledgements This work has been supported by the French government through the NIM-ML project (ANR-21-CE23-0005-01). All experiments were performed using the Julia 2 cluster, funded as DFG project as “Forschungsgroßgerät nach Art 91b GG” under INST 93/1145-1 FUGG.

References

1. Agarwal, C., Nguyen, A.: Explaining image classifiers by removing input features using generative models. In: Proceedings of the ACCV (2020)
2. Aghakishiyeva, G., Zhou, J., Arya, S., Poling, J.D., Houliston, H.R., Womble, J.N., Johnston, D.W., Bent, B.: Photorealistic inpainting for perturbation-based explanations in ecological monitoring. In: NeurIPS 2025 Workshop for Imageomics: Discovering Biological Knowledge from Images Using AI (2025)
3. Ang, D.: Daa233/generative-inpainting-pytorch: A pytorch reimplementation for paper generative image inpainting with contextual attention (February 2021), https://github.com/daa233/generative-inpainting-pytorch/commit/c6cdaea0427b37b5b38a3f48d4355abf9566c659, gitHub repository
4. Cheng, Z., Wu, Y., Li, Y., Cai, L., Ihnaini, B.: A comprehensive review of explainable artificial intelligence (xai) in computer vision. Sensors **25**(13) (2025)
5. Cheong, B.C.: Transparency and accountability in ai systems: safeguarding well-being in the age of algorithmic decision-making. Frontiers in Human Dynamics **6** (2024)
6. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. NeurIPS'20 (2020)
7. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. pp. 6970–6979. NeurIPS'17 (2017)
8. Elharrouss, O., Damseh, R., Belkacem, A.N., Badidi, E., Lakas, A.: Transformer-based image and video inpainting: current challenges and future directions. Artificial Intelligence Review **58**(4) (February 2025)
9. Garreau, D.: How to scale hyperparameters for quickshift image segmentation. In: Camps-Valls, G., Ruiz, F.J.R., Valera, I. (eds.) AISTATS 2022. Proceedings of Machine Learning Research, vol. 151, pp. 5243–5275 (28–30 Mar 2022)
10. Garreau, D., Mardaoui, D.: What does LIME really see in images? In: ICML 2021 (2021)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
12. Kendall, M.G., Smith, B.B.: The Problem of m Rankings. The Annals of Mathematical Statistics **10**(3), 275 – 287 (1939)
13. Kim, S., Suh, S., Lee, M.: RAD: Region-aware diffusion models for image inpainting. In: Proceedings of the CVPR. pp. 2439–2448 (June 2025)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the ICCV. pp. 4015–4026 (October 2023)
15. Kotelnikov, A., Baranchuk, D., Rubachev, I., Babenko, A.: TabDDPM: modelling tabular data with diffusion models. ICML'23 (2023)
16. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nature Communications **10**(1), 1096 (Mar 2019)
17. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. pp. 4768–4777. NIPS'17 (2017)
18. Mardaoui, D., Garreau, D.: An analysis of LIME for text data. In: Proceedings of The 24th AISTATS. Proceedings of Machine Learning Research, vol. 130, pp. 3493–3501 (13–15 Apr 2021)
19. mseitzer: Mseitzer/pytorch-fid: Compute fid scores with pytorch. (March 2024), https://github.com/mseitzer/pytorch-fid/commit/b9c18118d082cbd263c1b8963fc4221dc1cbb659

20. Papagiannidis, E., Mikalef, P., Conboy, K.: Responsible artificial intelligence governance: A review and research framework. The Journal of Strategic Information Systems **34**(2), 101885 (2025)
21. Petsiuk, V., Das, A., Saenko, K.: RISE: randomized input sampling for explanation of black-box models. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018. p. 151 (2018)
22. Radanliev, P.: AI ethics: Integrating transparency, fairness, and privacy in ai development. Applied Artificial Intelligence **39**(1), 2463722 (2025)
23. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144 (2016)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
25. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: 2022 WACV. pp. 3172–3182 (2022)
26. Suzuki, S., Abe, K.: Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing **30**(1), 32–46 (1985)
27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 CVPR. pp. 2818–2826 (2016)
28. Tan, Z., Tian, Y., Li, J.: Glime: General, stable and local lime explanation. In: Advances in Neural Information Processing Systems. vol. 36, pp. 36250–36277 (2023)
29. Tritscher, J., Lissmann, P., Wolf, M., Krause, A., Hotho, A., Schlör, D.: Generative inpainting for shapley-value-based anomaly explanation. In: Explainable Artificial Intelligence. pp. 230–243 (2024)
30. Vedaldi, A., Soatto, S.: Quick shift and kernel methods for mode seeking. In: Computer Vision – ECCV 2008. pp. 705–718 (2008)
31. Wang, Z., Simoncelli, E., Bovik, A.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers. vol. 2, pp. 1398–1402 Vol.2 (2003)
32. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. arXiv preprint arXiv:1806.03589 (2018)
33. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. arXiv preprint arXiv:1801.07892 (2018)
34. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision – ECCV 2014. pp. 818–833 (2014)
35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 CVPR. pp. 586–595 (2018)