Generative Inpainting for the Evaluation of Explanations*

Mohamed Jalal Baim[1], Romain Xu-Darme[2][0000-0002-8630-5635], and Alban Grastien[2][0000-0001-8466-8777]

[1] École Polytechnique, Institut Polytechnique de Paris, Palaiseau, France
mohamed.baim@polytechnique.edu
[2] Université Paris-Saclay, CEA, List, Palaiseau, France
{romain.xu-darme,alban.grastien}@cea.fr

Abstract. We present a novel method based on generative AI for evaluating the correctness of eXplainable AI (XAI) algorithms in the context of computer vision. Unlike current state-of-the-art methods that deactivate pixels by setting them to a fixed value (*e.g.* black) to confirm their contribution to the decision, our approach uses synthetic samples that are closer to both the original image and the original distribution of images. Using the Imagenette dataset, this approach leads to new empirical results when comparing the correctness of several XAI algorithms.

1 Introduction

In the context of computer vision, multiple algorithms from the field of eXplainable AI (XAI) aim to highlight which regions of the image (pixels) are *important* with respect to the decision, in the form of saliency maps. However, these algorithms do not always return the same explanation [1, 33]. When this happens, one would like to know which algorithm more correctly reflects the actual behavior of the model. Hence, several *metrics* have been proposed over the years to evaluate the *correctness* of explanation algorithms, by assuming that the model's decision will change if the important pixels are removed from its input.

Crucially, such metrics rely on the ability to efficiently *mask* (or *erase*) specific regions of the target image. While "erasing" a specific input feature in the context

* This work was supported by French government grants managed by the Agence Nationale de la Recherche under the France 2030 program with the references "ANR-23-DEGR-0001" and "ANR-23-PEIA-0006", as well as a Research and Innovation Action under the Horizon Europe Framework with grant agreement Nr.101070038. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This publication was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-De-France Regional Council. The code is available at https://github.com/Jalalbaim/GIEX-Generative-Inpainting-for-the-Evaluation-of-eXplanations.

of tabular data may be intuitive, removing a part of an image is not an easy task; it often boils down to setting all pixels to a fixed value [19, 5, 36] (*e.g.* black). However, evaluating an XAI algorithm using images with large regions of black pixels offers no guarantee as to the actual behavior of the model for the original image [7, 33, 13, 9, 4]. Indeed, such images are both out of the distribution of images seen during the training of the machine-learning model, and very different visually from the original image.

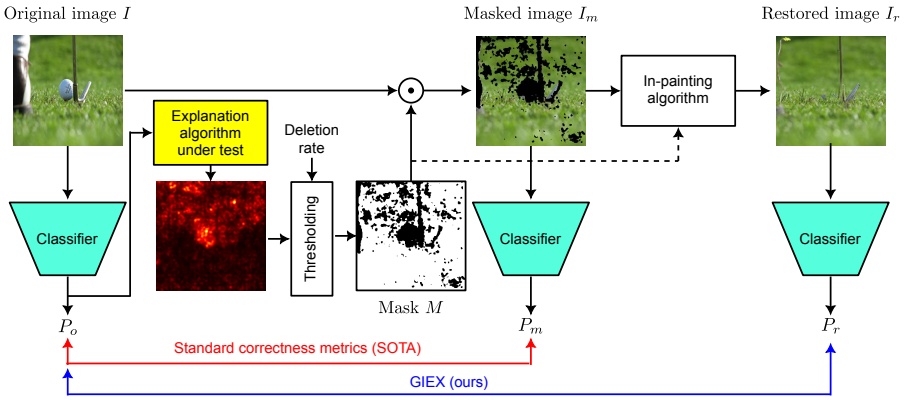

Fig. 1. GIEX pipeline overview. Rather than comparing the classification output of the original image to the output from a masked image, we first use an inpainting algorithm to cleanly remove the masked regions. P_o: probability of the target class on the original image; P_m: probability of the target class on the masked image; P_r probability probability of the target class on the image reconstructed by the inpainting algorithm. $\odot$ indicates an element-wise multiplication.

Thus, the purpose of the present work is to propose an evaluation metric that erases parts of the image in a more realistic manner: rather than simply replacing these parts with black pixels, we wish to cleanly remove them using contextual information (*e.g.* the neighborhood of the parts to erase). For example, as illustrated in Figure 1, if a classifier identifies a golf ball in an image, and an explanation algorithm indicates that the regions of the ball and the putter are important for the decision, we wish to remove both these objects from the image (leaving only a picture of grass) and to monitor the corresponding change in the model prediction. If the model no longer identifies the golf ball, then the explanation is correct – in the sense that it contains the most important regions of the image. Otherwise, the explanation does not reflect the behavior of the model and it is therefore incorrect.

Contributions In this context, we propose GIEX—for Generative Inpainting for the Evaluation of eXplanations—, a method for evaluating the correctness

of explanations. As with existing evaluation methods, the GIEX pipeline relies on the impact of removing explicative or non-explicative pixels on the model's decision; however, rather than simply replacing these pixels with arbitrary values, it uses generative AI to produce images that are in-distribution. With GIEX, we evaluate the explanations from seven XAI algorithms on different architectures, applied to images from the Imagenette dataset. Our results confirm that the respective quality of XAI algorithms depends on the architecture. They also contradict some results and, for instance, rehabilitate Guided Backpropagation as the best algorithm for VGG-16. A comprehensive study of out-of-distribution detection on the generated images demonstrates that the images produced during the pipeline are indeed closer to the original distribution than those used by existing methods, which adds weight to our method.

2 Related Work

Evaluation Metrics Most evaluation metrics manipulate the input by masking a percentage of the most important pixels and tracking the corresponding changes in the model's confidence. The goal is to test whether the pixels identified as important by an explanation algorithm truly drive the model's decision.

The Area Under the Deletion Curve [19] (**AUDC**) measures how quickly the model's confidence drops as important pixels are removed: lower AUDC values indicate that the explanation identifies truly important pixels. Conversely, the Area Under the Insertion Curve [19] (**AUIC**) captures how rapidly confidence increases when informative pixels are added to a baseline: higher AUIC values signify more correct explanations. Similarly, the **Average Drop (AD)** [5] and **Average Gain (AG)** [36] metrics summarize how the model's confidence increases/decreases when a fixed percentage of important pixels are removed. The ROAR approach [13] proposes to retrain a classifier on images where the most important pixels are removed, and to monitor the degradation in accuracy. Similarly, [9] propose to realign the training and testing data by introducing counterfactuals during the training process. Finally, the **Local Relative Correctness (LRC)** [33] measures how closely an explanation reflects the model's behavior in a small area around a given input, by comparing the base model to a local surrogate model built from the explanation.

Bringing masked images back into the distribution As mentioned above, most evaluation metrics that rely on masking important image regions produce samples that are out of distribution (OoD) – far from both the original image and the training distribution – which skews the evaluation. One possible method to bringing masked images back into the original distribution, proposed in the context of counterfactual generation [32], uses prototypical representations of each class as targets for the reconstruction of the image. Although effective on simple low-resolution data (*e.g.* MNIST [17]), scaling this method to higher resolution and colored images (*e.g.* CIFAR-10 [16], ImageNet [20]) remains challenging, which limits its practicality for the evaluation of explanation algorithms. In par-

ticular, after our own preliminary experiments with CIFAR-10, we concluded that such an approach generates high frequency noise (similar to adversarial attacks) rather than a semantically coherent transformation of the image.

Other related works [38, 35, 34] explore GAN-based methods. Contextual Attention Inpainting [35] uses a contextual attention layer to locate and copy relevant patches from known regions to fill missing areas. It follows a two-stage design (coarse-to-fine) and is trained with WGAN-GP [2, 8]. It delivers sharp, consistent textures but relies heavily on the availability of useful patches around the missing regions, producing unrealistic images when context is weak or irrelevant. The method also suffers from a high memory cost due to patch matching at high resolutions. SN-PatchGAN [34] introduces gated convolutions that learn soft gates per pixel and channel to select useful features. It uses a stable local discriminator – enabling effective training with just L_1 and GAN losses – and produces clean edges and natural colors. However, this architecture requires task-specific training and performs poorly when surrounding textures are misleading. Moreover, the method is still prone to hallucinations when guidance is weak. **DeepFillv2** [35] also builds on SN-PatchGAN by replacing standard gated convolutions with contextual feature transform layers, allowing the model to better capture long-range dependencies across the image. It also uses a coarse-to-fine architecture and trains with a combination of reconstruction and adversarial losses. The model produces more coherent completions, especially in large missing regions, and generalizes better across different mask shapes. However, it still struggles when the surrounding context is complex or ambiguous, and like other GAN-based methods, it can produce blurry or inconsistent textures in challenging cases. Structure-Guided GAN [38] first generates a clean edge map, then uses it to guide a texture generator that fills in colors and details. It combines gated convolutions, attention, and both local and global discriminators, optimizing with a mix of L_1, perceptual, style, and adversarial losses. This method generates sharper, more coherent results but it heavily relies on accurate edge prediction (poor edges lead to poor inpainting). Moreover, the model requires two generators, multiple discriminators, and careful tuning.

More recent works [18, 6] employ Denoising Diffusion Probabilistic Models (DDPMs), powerful likelihood-based generative models capable of producing high-quality, diverse samples by progressively denoising Gaussian noise into structured outputs. They operate in two phases: a forward process that gradually corrupts data with Gaussian noise, and a reverse process that learns to iteratively denoise, reconstructing the original data distribution. Once the generative model is trained, the image is split into two regions: known pixels, which remain unchanged, and unknown pixels, which are regenerated through an inpainting strategy. A naïve approach would be to condition the diffusion model directly on the known regions and to perform iterative denoising. Unfortunately, this type of approach often leads to edge artifacts, since smooth transitions between preserved and regenerated areas are hard to maintain. To leverage this issue, **RePaint** [18] introduces additional loops into the generative diffusion process to ensure homogeneity between kept and newly generated areas of an

image. Its only drawback *w.r.t.* GAN-based approaches is speed: diffusion sampling is slower than one GAN forward pass. LatentPaint [6] is another approach using DDPMs that introduces a latent-space inpainting method built on a pretrained diffusion model. It mixes known and missing regions directly within a UNet architecture at each level of information. As a result, inference is fast with no added computational cost, and it often outperforms RePaint in quality and speed. However, like all DDPM-based methods, it can hallucinate and its performance heavily depends on the base diffusion model, struggling more with complex scenes than with simpler ones (*e.g.* faces).

Note that in a work orthogonal to ours, [4] and [31] proposed to use generative AI as an inpainting strategy to generate explanations based on occlusion rather than to evaluate the correctness of explanations.

3 Evaluating explanations using generative inpainting

In this section, we present GIEX (Generative Inpainting for the Evaluation of eXplanations), our pipeline for the evaluation of XAI algorithms based on generative inpainting.

Pipeline overview GIEX evaluates explanations with perturbed samples close to the original data distribution. As illustrated in Figure 1, GIEX proceeds in three stages. First, given a test image I, the XAI algorithm under test computes a saliency map. In a second step, GIEX masks the top-K most relevant pixels of the image (or the least ones, depending on the protocol, see below) and uses an inpainting algorithm to fill in the missing part, thus producing a reconstructed image I_r that is in-distribution. At this point, the information from the masked pixels should have been removed from the image. Importantly, any inpainting algorithm could be swapped in at this stage, and we emphasize that *our goal is not to compare inpainting algorithms, but rather to show that inpainting yields correctness evaluation results that are different from the state of the art.* Finally, the classifier is run on the reconstructed image I_r, and the evaluation metrics are computed.

Deletion and Insertion protocols We evaluate each XAI algorithm under two complementary protocols:

Deletion. We progressively mask from 0 to 50% of the most important pixels (according to the XAI method under test), reconstruct each intermediate image I_r with an inpainting algorithm, and track the classifier's confidence P_r. Note that the 50% cap prevents inpainting large holes that may induce unrealistic hallucinations (see discussion in Section 4). If the explanation is correct, removing important pixels should induce a fast confidence drop (low AUDC; high AD / low AG).

Insertion. We start from an entirely black image and reveal pixels in order of importance (from 0 to 100%). For each insertion rate, the inpainting algorithm fills in the missing regions given the currently inserted set of pixels. A

correct explanation should induce a rapid confidence rise (high AUIC; high AG / low AD).

4 Experiments & Results

In this section, we present our experimental setup and results, with the goal of showcasing the limitations of state-of-the-art metrics described above, but also of studying the consensus between metrics from multiple perspectives.

4.1 Setup

Classifiers and dataset We conducted our experiments on two classifiers pretrained on the ImageNet [20] dataset: a ResNet-50 [10] and a VGG-16 [24] operating on 224×224 RGB inputs. The evaluation of the explanations was performed on the *Imagenette* [14] subset containing 10 classes, restricted to 15 images per class (due to the computational cost of RePaint).

Inpainting algorithm As indicated in Section 2, GAN-based algorithms can be sensitive to edge quality [38], or require task specific tuning [35], or struggle when the context is weak [34]. In contrast, pretrained DDPMs require no task-specific training, avoiding the adversarial loss tuning, large datasets, and domain-specific discriminators that GANs depend on. In practice, based on the lack of availability of the code for LatentPaint [6] at the time of this submission, we selected **RePaint** [18] and **DeepFillv2** [34] as our inpainting algorithms, as shown in Figure 2.

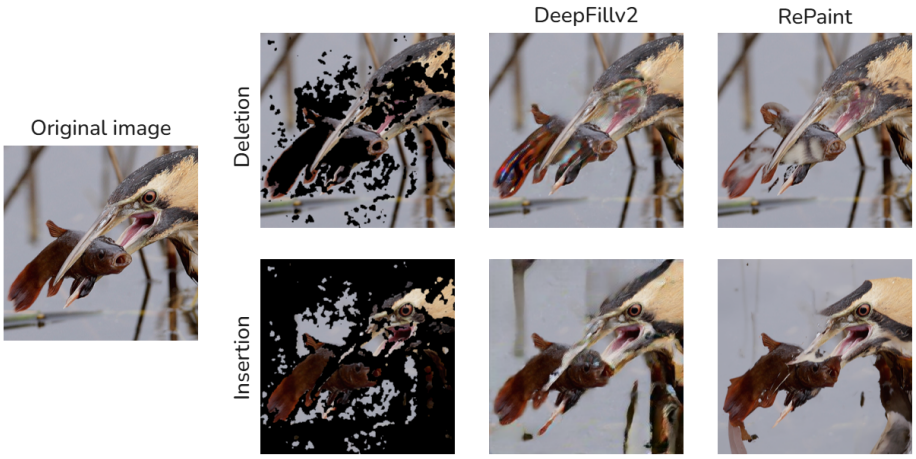

Fig. 2. Results of RePaint and DeepFillv2 for both approach. Deletion; 30% of pixels are masked. Insertion; 30% of pixels are revealed.

Explanation algorithms under test We evaluated the following popular XAI algorithms: Grad-CAM [21], Guided Backpropagation [28] (GBP), Saliency [23], Integrated Gradients [30] (50 steps), Input×Gradient [22], SmoothGrad [25] (50 samples) and LRP [3]. Given an attribution map, we aggregated pixel importance across all RGB channels by summing their absolute attributions, then applied a 5×5 Gaussian smoothing to mitigate the effects of pooling in the first layers of the classifier.

Assessing the quality of the reconstructed images Standard evaluation metrics for the correctness of XAI algorithms assume that the modified inputs remain meaningful for the model. However, simple inpainting with fixed-color fillings (*e.g.* black) drifts the image out of distribution. Thus, we borrow several metrics from the fields of out-of-distribution detection, used as proxies for the quality of the reconstructed images. In practice, we use the OpenOoD [37] framework, restricting ourselves to metrics that did not require an additional training step. For each XAI method and each metrics, we measure the "distance" between the original distribution of images $\mathcal{D}$ and the distribution of images with the 30% most important pixels removed $\mathcal{D}_r$ using the AUROC score and the False Positive Rate at 95% True Positive Rate (FPR@95): an AUROC score close to 50 and a high FPR@95 score indicate that the distributions are indistinguishable, *i.e.* that images in $\mathcal{D}_r$ remain in-distribution; a high (or low) AUROC score and a small FPR@95 indicate that the distributions can be separated, *i.e.* that images in $\mathcal{D}_r$ are out-of-distribution. As an additional test, we also computed the Fréchet Inception Distance (FID [12]) metric between $\mathcal{D}$ and $\mathcal{D}_r$.

Evaluation metrics We reported AD [5] and AG [36] at a fixed deletion rate (20%), on masked (black) images (standard approach) and after inpainting with RePaint and Deepfillv2. As a baseline, we also blurred the region to be masked. For Deletion, we computed the AUDC [19] on black-masked images, and its counterpart, the Area Under the Repainted Curve (AURC), obtained after inpainting. Finally, we also computed the LRC [33] to explore local correctness under small perturbations. Note that we exclude ROAR from our setup, due to its prohibitive cost in terms of computation (one model training per XAI algorithm per deletion rate).

4.2 Results

Does the GIEX pipeline evaluate XAI algorithms differently than the state of the art?

In the Deletion protocol, removing the most relevant pixels should lead to a noticeable drop in confidence ($P_r/P_m \ll P_o$). Therefore, a high AD and a low AG indicate a correct explanation algorithm. Similarly, both AUDC and AURC should remain low, meaning that the model quickly loses confidence when the most informative pixels are masked.

Using ResNet-50 as the base classifier (see Table 1), the best-performing algorithm in the standard setting (SOTA) is GBP, which achieves the lowest

Methods	SOTA				GIEX (RePaint)			GIEX (DeepFillv2)		
	$AUDC \downarrow$	$AG_{20} \downarrow$	$AD_{20} \uparrow$	$LRC \downarrow$	$AURC \downarrow$	$AG_{20} \downarrow$	$AD_{20} \uparrow$	$AURC \downarrow$	$AG_{20} \downarrow$	$AD_{20} \uparrow$
ResNet50										
GBP	**3.52**	**0.03**	25.25	71.57	13.33	5.21	32.16	10.98	5.73	46.34
GradCam	17.23	5.58	**35.35**	869.92	14.34	3.65	33.16	11.84	**2.34**	30.22
Saliency	10.77	2.78	16.16	242.07	17.37	3.80	17.24	13.06	7.10	18.08
Int.Grad	11.68	4.76	20.20	272.05	16.30	4.27	23.48	12.72	5.39	21.87
SmoothGrad	7.46	2.07	9.60	268.52	**10.57**	**3.36**	**41.79**	**10.76**	4.23	**47.73**
Input×Grad	13.92	5.71	32.32	207.13	17.69	4.48	16.01	13.13	6.43	15.81
LRP	15.03	7.64	31.82	**56.07**	18.56	4.27	14.27	13.16	4.57	13.90
VGG-16										
GBP	**4.13**	3.37	43.69	**5.80**	22.07	**2.71**	**96.14**	**17.91**	**2.08**	**56.94**
GradCam	16.54	**2.69**	30.59	81.55	27.57	5.52	64.51	21.58	4.51	26.40
Saliency	8.11	4.22	33.25	10.89	24.91	4.37	85.88	20.91	2.82	33.19
Int.Grad	8.31	3.25	38.53	37.19	23.40	8.31	84.37	19.95	2.96	40.92
SmoothGrad	10.23	3.97	36.14	20.79	**18.44**	6.78	81.29	18.47	2.57	50.15
Input×Grad	10.47	5.15	30.93	15.07	26.51	5.36	78.34	21.13	2.32	30.76
LRP	10.61	7.91	**77.73**	13.04	27.54	4.78	63.96	20.23	2.62	31.17

Table 1. Comparison of metric results (SoTA at 20%, GIEX at 20%). AUDC, AURC, and LRC values represent the average value computed over the entire dataset. ↓: lower is better. ↑ higher is better.

AUDC and AG scores, and good AD and LRC scores. However, using our GIEX pipeline, we determine that SmoothGrad achieves the best results across all metrics (AURC, AD and AG) when using RePaint and the best AURC/AD when using DeepfillV2.

Using VGG-16 as the base classifier, GBP once again stands out as the best-performing algorithm in the standard configuration (SOTA), achieving the lowest AUDC and LRC. Using our GIEX pipeline, SmoothGrad achieves the best AURC score under RePaint inpainting, while GBP consistently yields the lowest AG and highest AD scores across all inpainting methods. To sum up, GBP consistently seems to deliver the most correct explanations for VGG-16, according to both standard metrics (SOTA) and our GIEX metrics. Interestingly, this result directly contradicts some criticisms [1] regarding the correctness of GBP, and reinforces the claims of subsequent works [11].

Overall, the comparison highlights that while both networks – ResNet-50 and VGG-16 – exhibit slightly similar metric rankings, VGG-16 seems relatively more stable (smaller output variability for similar images) than ResNet-50, as seen by its lower LRC values across all XAI algorithms. This result shows that VGG-16's feature representations are less sensitive to small perturbations in the input, and more generally suggests that the architecture itself influences explanations: models like VGG-16, with simpler and more sequential feature extraction, favor globally coherent and stable explanations, while deeper architectures such as ResNet-50 produce more sensitive explanations. This highlights the importance of considering architectural effects when comparing XAI algorithms across models.

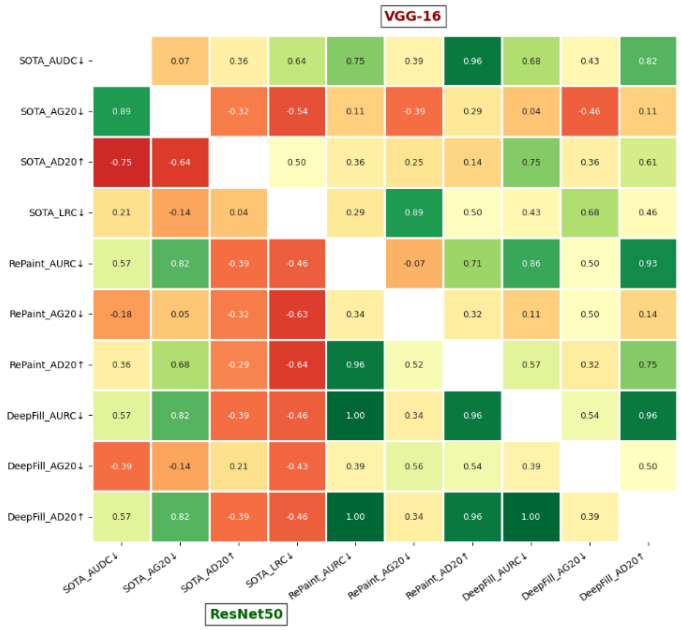

Fig. 3. Pairwise Spearman Rank Correlation Scores between evaluation metrics for the correctness of XAI algorithms. A score of 1.0 (resp. -1.0) indicates that the two metrics rank the XAI algorithms from most to least correct in the same (resp. opposite) order.

To go beyond the selection of the most correct explanation algorithm, we measured the pairwise consensus between evaluation metrics as their Spearman Rank Correlation Score, *i.e.* two evaluation metrics agree if they rank the XAI algorithms from most to least correct in the same order [33]. From the correlation matrices in Figure 3, it can be observed that evaluation metrics do not always agree on the most correct explanation algorithms.

For VGG-16, the standard deletion-based metrics (AD, AG – SOTA) exhibit strong mutual correlations, indicating that they consistently measure a similar behavior across explanation algorithms. Likewise, the RePaint-based variants obtained with our GIEX pipeline (AD, AG – ours) show a high degree of internal coherence, confirming that the RePaint evaluation remains stable within its own configuration. However, correlations between standard and RePaint metrics remain weak or even negative, suggesting that both families capture distinct aspects of correctness and model sensitivity. Global metrics such as AURC, AUDC, and LRC appear poorly aligned with either family, highlighting their limited agreement with this type of measures.

For ResNet-50, a similar trend is observed but with slightly lower overall correlations. The standard metrics still form a coherent group, while RePaint metrics correlate mainly among themselves. Cross-group correlations are weaker than for VGG-16, indicating that architectural differences further affect the re-

lationship between metrics. As for VGG-16, AURC, AUDC, and LRC remain mostly uncorrelated with other metrics.

Overall, both matrices confirm that *no consistent consensus exists across evaluation metrics*. While standard and RePaint measures remain internally coherent, they diverge sharply from each other and from the global metrics, demonstrating that each family reflects a different notion of explanation correctness.

Are the results similar when using different inpainting algorithms? The results in Figure 3 show that there is no consensus between SOTA and inpainting-based methods, as their cross-method correlations are weak to moderate on both architectures. However, RePaint and DeepFill show strong mutual consistency, with several metric pairs. This indicates that conclusions drawn from one inpainting method generalize reliably to the other method.

Are the results stable? The RePaint algorithm is stochastic, which could impact the consistency of our conclusions. To evaluate this impact, we ranked all XAI methods using our AURC metric with 4 different seeds, and we measured a $[0.94, 1.0]$ spearman rank correlation between these runs, indicating that the rankings obtained are indeed stable.

Are the reconstructed images in-distribution? As shown in Table 2 and 3, for the vast majority of OoD detection metrics, the distribution of images reconstructed using an inpainting algorithm is closer to the original distribution than masked images or blurred images (AUROC score closer to 50, high FPR@95). We obtain similar results for explanations generated by Integrated Gradients, GradCAM, Saliency, LRP and Input×Gradient and on VGG-16 (see Supplementary material). Interestingly, the results differ slightly for GradNorm [15], RankFeat [27] and KNN [29] but we argue that the presence of these outliers reflect a potential flaw in the corresponding metrics rather than a questioning of our results. These results demonstrate that inpainting effectively mitigates the distribution shift problem, confirming our hypothesis and allowing for a more stable and trustworthy evaluation of XAI algorithms. As an additional experiment, we report the average L_2 distance and the FID [12] (perceptual similarity) between the original image and the masked/inpainted image in Table 4 and Table 5 respectively. This experiment highlights a discrepancy between the distance in the input space (L_2 distance) – where blurred images are consistently closer to the original image in the input space – and the results using OoD detection metrics and perceptual similarity (FID) – where inpainted images are more visually similar to the original images. Nevertheless, in all cases, images with black masks are systematically the worst alternative to evaluate XAI methods and should not be used.

Hallucinations and limitations While GIEX indeed shows clear advantages in maintaining the images within the original data distribution, algorithms using generative diffusion models, such as RePaint, can be prone to hallucinations at higher deletion rates, or to produce semantically inconsistent regions that could

Method	RePaint	DeepFill	Blurred	Masked
ash	**77.35**/**74.00**	86.70/46.00	90.48/40.67	86.21/44.00
dice	**69.72**/**74.00**	80.77/58.00	88.64/34.67	83.68/44.67
ebo	**77.35**/**74.00**	86.70/46.00	90.48/40.67	86.21/44.00
fdbd	**79.40**/**70.67**	86.29/36.00	93.82/26.67	95.18/22.67
gen	**77.54**/**78.00**	86.64/50.00	89.80/49.33	86.40/44.67
gradnorm	**62.08**/**89.33**	77.40/66.67	81.88/56.00	69.32/77.33
klm	**85.87**/**70.67**	94.05/48.00	99.46/0.00	99.97/0.00
knn	**45.96**/**94.00**	45.15/**94.00**	71.54/76.67	76.72/58.67
mls	**77.49**/**74.00**	86.69/45.33	90.00/42.00	86.09/42.00
msp	**74.92**/**84.67**	83.16/51.33	85.94/49.33	82.49/49.33
nci	**74.55**/**68.67**	85.89/50.00	90.89/25.33	86.93/34.67
odin	**69.44**/**76.67**	74.36/66.00	82.00/58.00	88.12/40.00
rankfeat	**56.08**/**96.00**	60.07/92.67	62.99/90.67	62.64/90.00
vim	**79.87**/**58.00**	83.97/43.33	95.64/18.67	99.84/1.33
vra	**79.74**/**71.33**	89.06/37.33	93.96/25.33	93.74/23.33

Table 2. AUROC/FPR@95 scores between the distributions of original images and images with most important pixels removed using Guided-propagation on ResNet-50, using popular OoD detection metrics. AUROC closer to 50 is better, higher FPR@95 is better.

Method	RePaint	DeepFill	Blurred	Masked
ash	**83.22**/**58.00**	84.55/57.33	89.05/44.00	91.51/35.33
dice	**73.89**/**70.00**	78.94/68.67	80.24/62.00	74.72/56.00
ebo	**83.22**/**58.00**	84.55/57.33	89.05/44.00	91.51/35.33
fdbd	84.17/56.00	**83.77**/**68.67**	90.56/34.67	95.80/21.33
gen	**83.14**/60.67	84.26/**66.00**	89.20/37.33	91.70/32.00
gradnorm	65.72/**96.00**	74.66/78.00	75.75/80.00	**64.58**/86.00
klm	91.84/65.33	**88.64**/**68.67**	94.74/56.67	98.38/7.33
knn	36.59/**98.00**	**42.97**/97.33	37.71/97.33	33.64/96.67
mls	**83.18**/**60.00**	84.54/58.00	89.24/41.33	91.63/33.33
msp	**79.91**/65.33	81.37/**70.67**	86.71/51.33	89.45/38.67
nci	**78.63**/**68.00**	81.77/57.33	84.46/57.33	82.57/48.00
odin	78.82/**75.33**	**77.87**/69.33	86.05/60.00	91.78/36.67
rankfeat	54.59/91.33	60.28/86.00	58.58/**92.67**	**53.82**/90.67
vim	83.31/58.00	**76.32**/**60.00**	85.47/51.33	97.94/10.67
vra	**85.10**/46.67	87.56/**52.00**	91.00/42.67	93.54/21.33

Table 3. AUROC/FPR@95 scores between the distributions of original images and images with most important pixels removed using Smoothgrads on ResNet-50, using popular OoD detection metrics. AUROC closer to 50 is better, higher FPR@95 is better.

Methods	@20%				@40%			
	Black	Blurred	RePaint	DeepFill	Black	Blurred	RePaint	DeepFill
Input×Grad	27116	5449	6030	**4012**	38277	**8334**	13464	8687
Int.Grad	27104	6162	6537	**4734**	38061	**9083**	13474	9513
LRP	24856	7120	8063	**6778**	35755	**9862**	13463	11001
GradCam	25151	**6923**	10883	8908	36605	**9405**	17315	13604
Saliency	25028	4753	4534	**3482**	36105	7473	10433	**7411**
GBP	24943	8818	9304	**8360**	35435	**11358**	15021	13140
SmoothGrad	25615	**8429**	12087	10967	36127	**11041**	18999	16768

Table 4. Average L_2 distance between the original image and the black/blurred/inpainted images with $X\%$ of most important pixels removed for ResNet-50. Lower is better.

Method	**Masked**	**Blurred**	**GIEX(Repaint)**	**GIEX(Deepfillv2)**
GBP	437.22	375.54	**330.45**	341.67
GradCam	427.94	250.92	240.69	**185.11**
InputxGrad	395.78	234.77	**167.25**	173.73
Int.Grad	476.66	309.92	**229.49**	249.08
Saliency	437.91	295.39	214.89	**160.89**
SmoothGrad	443.71	339.96	307.45	**281.71**
LRP	395.78	234.77	201.95	**173.73**

Table 5. FID ↓ comparison between generated images per method

change the model's confidence in ways unrelated to the original content (see Figure 4). To this, we argue that: i) because we analyse the output logits rather than the output of a softmax function, when important features for class A are replaced by a hallucination (*e.g.* class B), then the decrease in the logit of class A is due to the feature removal rather than the increase of the logit of class B; ii) the correctness of an explanation is a *local* property (*i.e.* corresponding to relatively low deletion rates). Note that conversely, for *very* low deletion rates (*e.g.* lower than 10%), inpainting tends to reconstruct the original image, which defeats our purpose. Hence, the effective range of deletion of our method is between 10% and 50%, which is sufficient to capture differences in correctness between the different XAI methods under test, as shown above.

5 Conclusion

In this work, we introduced GIEX, a new evaluation framework that uses generative inpainting through diffusion models to assess the correctness of XAI algorithms in computer vision. We demonstrated through experiments that this strategy is able to evaluate the correctness of explanations using images that remain closer to the original distribution than the state-of-the-art approaches. The analysis performed on both VGG-16 and ResNet-50 architectures clearly

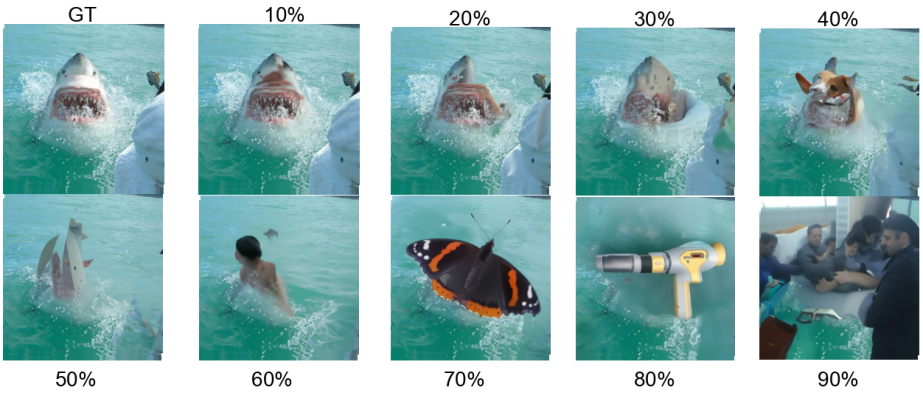

Fig. 4. Examples of hallucinations from RePaint at high deletion rates.

exhibits a lack of consensus between metrics on which explanation algorithm is the most correct. More importantly, such results seem architecture-dependent, indicating that the evaluation of the correctness of XAI algorithms must be performed for each new use-case.

While GIEX indeed shows clear advantages in maintaining the images within the original data distribution, we have also identified some avenues of research. Firstly, the use of generative AI in our approach comes with an additional computational cost, compared to state-of-the-art evaluation metrics that use a black inpainting. As an alternative to DDPMs, Denoising Diffusion Implicit Models [26] (DDIMs) use a deterministic – and therefore faster – generation process that could be applied to our work, in an effort to mitigate this computational overhead. More generally, it could be interesting to compare the results obtained using more inpainting algorithms [38, 6, 35]. Secondly, while this work focused on image classification, our GIEX evaluation pipeline is highly generic and could be applied to evaluate the correctness of XAI algorithms for different modalities (*e.g.* tabular data, text) and different tasks (*e.g.* regression, segmentation).

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: International Conference on Neural Information Processing Systems (NeurIPS-18). pp. 9525—-9536 (2018)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning (ICML-2017). pp. 214–223 (2017)
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**(7) (2015)
4. Chang, C.H.K., Creager, E., Goldenberg, A., Duvenaud, D.K.: Explaining image classifiers by counterfactual generation. In: International Conference on Learning Representations (2018)

5. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV-18). pp. 839–847 (2018)
6. Corneanu, C.A., Gadde, R., Martínez, A.M.: LatentPaint: Image inpainting in latent space with diffusion models. In: IEEE Winter Conference on Applications of Computer Vision (WACV-24). pp. 4322–4331 (2024)
7. Gomez, T., Fréour, T., Mouchère, H.: Metrics for saliency map evaluation of deep learning explanation methods. In: International Conferences on Pattern Recognition and Artificial Intelligence (ICPRAI-22). pp. 84–95 (2022)
8. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: International Conference Neural Information Processing Systems (NeurIPS-17). pp. 5767–5777 (2017)
9. Hase, P., Xie, H., Bansal, M.: The out-of-distribution problem in explainability and search methods for feature importance explanations. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2021)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR-16). pp. 770–778 (2016)
11. Hedström, A., Weber, L., Lapuschkin, S., Höhne, M.M.C.: A fresh look at sanity checks for saliency maps. In: World Conference on Explainable Artificial Intelligence (XAI-24). pp. 403–420 (2024)
12. 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: Neural Information Processing Systems (2017)
13. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. In: International Conference on Neural Information Processing Systems (NeurIPS-18). pp. 9734–9745 (2018)
14. Howard, J., et al.: Imagenette. https://github.com/fastai/imagenette (2020), accessed 22/06/2026
15. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. In: International Conference on Neural Information Processing Systems (NeurIPS-21). pp. 677–689 (2021)
16. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009)
17. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**, 2278–2324 (1998)
18. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Gool, L.V.: RePaint: Inpainting using denoising diffusion probabilistic models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR-22). pp. 11451–11461 (2022)
19. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized input sampling for explanation of black-box models. In: British Machine Vision Conference (BMVC-18) (2018)
20. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**, 211–252 (2014)
21. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**, 336–359 (2016)

22. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning (ICML-17). pp. 3145–3153 (2017)
23. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Workshop at International Conference on Learning Representations (2014)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR-15) (2015)
25. Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise. Proceedings of the ICML Workshop on Visualization for Deep Learning (2017)
26. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR-21) (2021)
27. Song, Y., Sebe, N., Wang, W.: RankFeat: Rank-1 feature removal for out-of-distribution detection. In: International Conference on Neural Information Processing Systems (NeurIPS-22) (2022)
28. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for simplicity: The all convolutional net. In: Workshop at International Conference on Learning Representations (2014)
29. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: International Conference on Machine Learning (ICML-2022). pp. 20827–20840 (2022)
30. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning (ICML-17). pp. 3319–3328 (2017)
31. Tritscher, J., Lissmann, P., Wolf, M., Krause, A., Hotho, A., Schlör, D.: Generative inpainting for Shapley-value-based anomaly explanation. In: World Conference on Explainable Artificial Intelligence (XAI-24) (2024)
32. Van Looveren, A., Klaise, J.: Interpretable counterfactual explanations guided by prototypes. In: European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD-21). p. 650–665 (2021)
33. Xu-Darme, R., Benois-Pineau, J., Giot, R., Quenot, G., Chihani, Z., Rousset, M.C., Zhukov, A.: On the stability, correctness and plausibility of visual explanation methods based on feature importance. In: International Conference on Content-based Multimedia Indexing (CBMI-2023). pp. 119–125 (2023)
34. Yu, J., Lin, Z.L., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. IEEE International Conference on Computer Vision (ICCV-19) pp. 4470–4479 (2018)
35. Yu, J., Lin, Z.L., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR-18). pp. 5505–5514 (2018)
36. Zhang, H., Torres, F., Sicre, R., Avrithis, Y., Ayache, S.: Opti-CAM: Optimizing saliency maps for interpretability. Computer Vision and Image Understanding **248**, 104101 (2024)
37. Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., Li, Y., Liu, Z., Chen, Y., Li, H.H.: OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. ArXiv (2023)
38. Zhao, L.J., Zhu, T., Wang, C., Tian, F., Yao, H.: Image inpainting algorithm based on structure-guided generative adversarial network. Mathematics **13**(15), 2370 (2025)