# Self-Guided Integrated Gradient Method for Attribution

Sabrina Henry* Alice Ruget Stirling Scholes Jonathan Leach
Heriot-Watt University, Edinburgh, UK
*sjh9@hw.ac.uk

## Abstract

*Explaining the decisions of deep neural networks is essential for building trust in AI systems, particularly when deployed in sensitive domains such as healthcare, security, and autonomous transport. Path-based attribution methods such as Integrated Gradients provide local explanations by integrating model gradients along a path from a baseline to the input image. However, such methods require user-defined baselines and often accumulate noisy gradients from saturated regions of the prediction landscape. To address these limitations, we propose the **S**elf-guided **I**ntegrated **G**radient **M**ethod for **A**ttribution (SIGMA), a baseline-free path-based method that stochastically explores the model's confidence landscape to identify input features responsible for the collapse of class confidence. By following the model's decision boundary, SIGMA produces interpretable and reliable attributions without requiring reference inputs or access to a model's internal representation layers. Evaluations across diverse architectures, including Vision Transformers, and computer vision datasets in healthcare and security, demonstrate that SIGMA provides spatially coherent attribution maps with strong faithfulness to the model's internal reasoning. Additionally, SIGMA generates zero-confidence variants of the input, recognisable to humans but adversarial to the model. We show that augmenting the original training dataset and retraining with these samples improves both robustness to noise and resistance to adversarial attacks. The code is publicly available at:* https://github.com/HWQuantum/SIGMA

## 1. Introduction

With machine learning systems becoming increasingly adopted in critical sectors such as healthcare [1–4], finance [5–7], security [8–10], and autonomous vehicles [11–14], the need for interpretability has become essential to building user trust [15]. At the same time, the growing complexity of these models makes their internal reasoning difficult to understand [16]. In response, research in eXplainable Artificial Intelligence (XAI) has emerged to develop methods that help humans interpret, and when appropriate, trust, the decisions of modern AI systems [17].

Feature attribution methods provide post-hoc, local explanations by estimating how each input feature contributes to a model's prediction [18]. Broadly, these methods can be grouped into perturbation-based and gradient-based approaches. Perturbation-based techniques, such as occlusion [19, 20], LIME [21], and SHAP [22], assess feature importance by masking or perturbing regions of the input and measuring the corresponding change in model output. While intuitive, such methods are often sensitive to the choice of masking strategy, with resulting attributions that may lack spatial precision [23]. In contrast, gradient-based methods leverage the model's internal derivatives to infer feature relevance, allowing for pixel-level attributions [24].

Among these gradient methods, path-based techniques such as Integrated Gradients (IG) [25] assign feature importance by integrating the gradient of the model output with respect to its input features along a continuous path in input space. In IG, this path is a straight line interpolating between a baseline, a reference input representing the absence of information (such as a black image), and the input being explained. Subsequent work has shown that IG attributions are sensitive to the choice of baseline [26] and can accumulate noise from irrelevant gradients [27]. Several extensions have been proposed to mitigate the noise in attributions [28, 29], with Guided Integrated Gradients (GIG) [30] introducing the idea of an adaptive path. Rather than perturbing all features uniformly, GIG adaptively updates only those with the smallest partial derivatives, producing less noisy attributions. While effective, GIG and other IG variants still rely on user-defined baselines, introducing subjectivity and instability into the resulting explanations [26].

Extending this line of work, IG² [31] formulates attribution as a counterfactual problem. It constructs a path by minimising the distance in representation space between the input to be explained and a manually chosen counterfactual reference. Gradients from the input and counterfactual classes are then combined during integration to produce an attribution map. This approach depends on the selection of

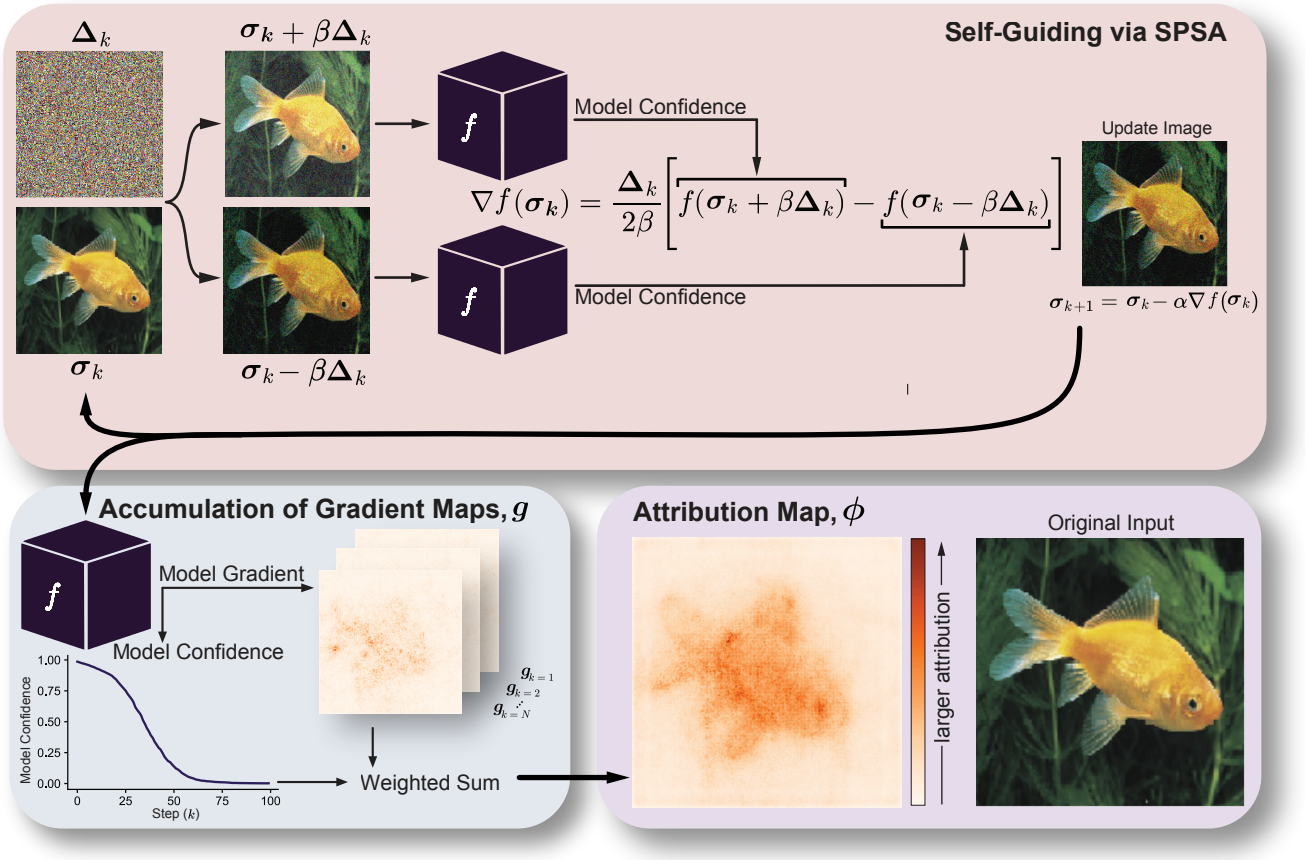


Figure 1. Illustration of the proposed SIGMA path method. The input image, $\sigma_{k=0}$ is iteratively perturbed to progressively reduce the model's confidence in the target class $f$ toward zero. At each iteration, the pixel-wise gradients of the softmax confidence with respect to the input image are computed and accumulated, resulting in a cumulative attribution map $\phi$ highlighting influential regions to the prediction of the target class.

a counterfactual reference, with the resulting attribution being sensitive to different types of reference [31]. It also requires access to intermediate model representations. Selecting such representation layers requires architectural knowledge and may be non-trivial for modern architectures such as Vision Transformers (ViTs), whose internal representations differ from traditional CNNs [32].

To address these challenges, we introduce the Self-guided Integrated Gradient Method for Attribution (SIGMA), a baseline-free method that constructs its path directly from the model's confidence landscape. Rather than relying on a manually defined baseline, counterfactual reference or access to internal representation layers, SIGMA stochastically explores the input space using the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm [33]. At each iteration, SIGMA perturbs the input in random directions and estimates a descent direction that drives the model's confidence in the predicted class toward zero. Gradients are accumulated along this self-guided path and weighted by the corresponding drop in confidence, ensuring sensitivity to informative gradients. See Figure 1 for an illustration of the algorithm with further implementation details discussed in Section 3.1. We summarise our contributions as follows:

- We propose SIGMA, a path-based attribution method that stochastically explores a model's confidence landscape to construct attribution paths that avoid regions of gradient saturation.
- SIGMA eliminates the need for an arbitrary baseline, counterfactual class selection, or internal model representation layers.
- We evaluate SIGMA across diverse network architectures and computer vision tasks, including real-world applications in healthcare and security, and demonstrate its extension to sub-patch interpretability in Vision Transformers.
- The stochastic nature of SIGMA supports path averaging and enables the computation of confidence metrics, pro-

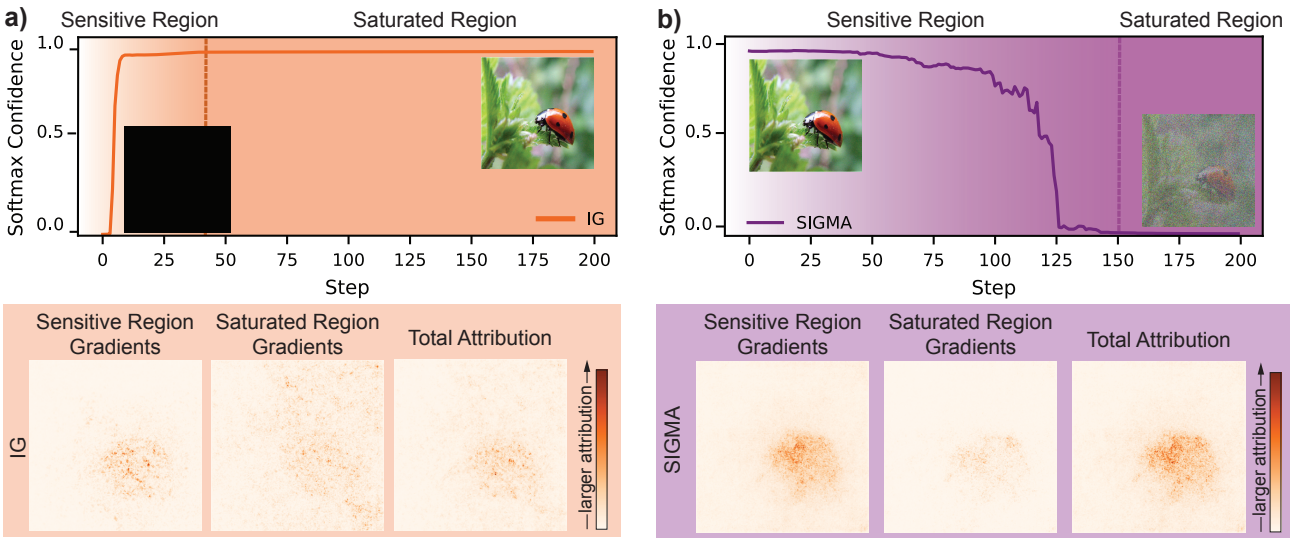


Figure 2. Comparison of attribution paths for Integrated Gradients (IG) and SIGMA for the input image shown inset and the class 'Ladybug'. (a) The IG path rapidly reaches a high-confidence plateau where gradients become saturated and uninformative. (b) SIGMA explores the model's confidence landscape, progressively reducing the model's confidence without saturation. Gradient visualisations from the sensitive and saturated regions, as well as the total attribution, are shown below each plot. IG accumulates noise from the saturated region, whereas SIGMA maintains spatially coherent, meaningful gradients throughout. This saturation behaviour is also observed when using logits rather than probabilities, but we report softmax confidence for consistency with XAI literature [27].

viding a quantitative measure of attribution map reliability.

- SIGMA generates noisy augmentations of the original data, which are zero-confidence variants of the input that exploit the model's decision boundary. These images remain recognisable to humans but receive a near-zero confidence from the network, mimicking an adversarial attack. Incorporating these samples during retraining can improve model robustness to noise and increase resistance to adversarial attacks.

## 2. Accumulation of Gradients

The goal of an attribution method is to identify which input features most influence the model's confidence. In computer vision this involves revealing the pixels that the network is most sensitive to when forming its prediction. IG achieves this by integrating the model gradient along a straight path from a baseline $\boldsymbol{x'}$ to the input image $\boldsymbol{x}$:

$$\phi_i^{IG}(f, \boldsymbol{x}, \boldsymbol{x'}) = (x_i - x_i') \int_0^1 \frac{\partial f(\boldsymbol{x'} + \alpha(\boldsymbol{x} - \boldsymbol{x'}))}{\partial x_i} \, d\alpha \tag{1}$$

Here, $i$ indexes the input feature (or pixel), and each gradient is weighted by the total change of the corresponding input feature from the baseline, $(x_i - x_i')$, which arises directly from the path integral formulation. While this satisfies the desirable completeness property, ensuring that the sum of the attribution equals the change in model output [25], it does not control how individual gradients along the path contribute to the final attribution. As a result, gradients computed in regions where the model's confidence is saturated can accumulate in the final attribution, even though they have little influence on the prediction. Figure 2(a) illustrates this behaviour by showing the model's confidence along the IG path, highlighting a sensitive region where confidence changes rapidly and a saturated region where it plateaus, a common observation in IG [26, 27, 30]. Gradients accumulated in the saturated region introduce noise that overshadows the more meaningful contributions from the sensitive region, causing the final attribution to inherit this noise.

In contrast, SIGMA stochastically explores the model's confidence landscape without requiring a baseline. The input is perturbed in random directions, and incremental updates are taken along the direction that minimises model confidence in the predicted class. This process avoids the early saturation observed in IG and instead constructs a gradual trajectory that captures multiple stages of confidence decay, resulting in a more complete and informative attribution. This behaviour is visualised in Figure. 2(b), where SIGMA continues to accumulate informative gradients even in regions where IG becomes saturated.

To assess the contribution of gradients accumulated along the path, we use the Softmax Information Curve (SIC) [34], shown in Figure 3. The SIC quantifies how effectively an attribution map reconstructs the model's con-

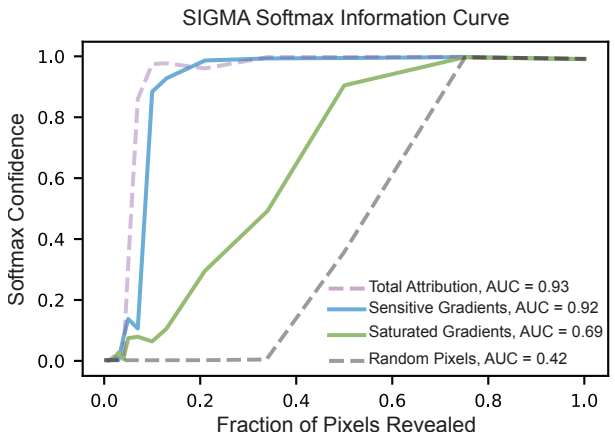


Figure 3. Softmax Information Curve (SIC) analysis of gradients along the SIGMA path for the same input shown in Fig. 2(b). Gradients accumulated from saturated regions of the confidence curve contribute less to reconstructing the model's confidence than those from sensitive regions, where there are larger changes in confidence. Nevertheless, they still encode useful information about the model's internal response, as indicated by a higher AUC compared to revealing pixels in random order.

fidence as image regions are progressively revealed in order of attribution importance. An ideal method recovers confidence rapidly, producing a high area under the curve (SIC-AUC). As seen in Figure 3, gradients from saturated regions contribute less to this reconstruction than those from the sensitive region, indicating that uniform accumulation can overemphasize less informative areas. To address this, SIGMA weights each gradient map by the model's change in confidence between successive steps, ensuring that each step's attribution preserves the relative influence of gradients while maintaining the completeness property.

## 3. The Self-Guided Integrated Gradient Method for Attribution

### 3.1. Path Construction

The core of the SIGMA attribution method is summarised in Figure 1. The first block illustrates the path construction inspired by SPSA [33], an optimisation algorithm widely adopted as an alternative to traditional gradient descent for problems with noisy or high-dimensional cost landscapes [35–37], making it particularly well suited for exploring the complex prediction landscape of deep neural networks. At each iteration $k$, the current parameter vector $\boldsymbol{\sigma}_k$ is perturbed simultaneously along all dimensions using a random perturbation vector $\boldsymbol{\Delta}_k$, whose components are drawn from a symmetric Bernoulli distribution ($\pm 1$ with equal probability). Two evaluations of the objective function are then performed at symmetrically perturbed points:

$$f(\boldsymbol{\sigma}_k + \beta\boldsymbol{\Delta}_k) \quad \text{and} \quad f(\boldsymbol{\sigma}_k - \beta\boldsymbol{\Delta}_k),$$

where $\beta$ controls the magnitude of the perturbation. The stochastic gradient estimate of the objective function $f(\boldsymbol{\sigma}_k)$ is obtained from these two measurements as

$$\nabla f(\boldsymbol{\sigma}_k) = \frac{f(\boldsymbol{\sigma}_k + \beta\boldsymbol{\Delta}_k) - f(\boldsymbol{\sigma}_k - \beta\boldsymbol{\Delta}_k)}{2\beta}\,\boldsymbol{\Delta}_k, \qquad (2)$$

This estimated gradient is then used to update $\boldsymbol{\sigma}_k$ in the descent direction:

$$\boldsymbol{\sigma}_{k+1} = \boldsymbol{\sigma}_k - \alpha\nabla f(\boldsymbol{\sigma}_k), \qquad (3)$$

where $\alpha$ is the step-size parameter. Through this simultaneous perturbation and two-sample gradient estimate, SPSA efficiently approximates the descent direction with only two function evaluations per iteration, regardless of the dimensionality of $\boldsymbol{\sigma}$.

In the context of SIGMA, $f(\boldsymbol{\sigma})$ represents the model's prediction landscape for the originally predicted (target) class, defined by the model's confidence score $f(\boldsymbol{\sigma}_k)$ as a function of the input image $\boldsymbol{\sigma}_k$, with $\boldsymbol{\sigma}_{k=0}$ the original input and $\boldsymbol{\sigma}_{1\leq k\leq N}$ each perturbed image along the path. Here, $f$ denotes the model's output corresponding to the target class, this may be logits or normalised softmax probability, for the purpose of this work we refer to model confidence as softmax probability. As shown in Figure 1 the random perturbation $\boldsymbol{\Delta}_k$ is applied independently to each pixel of the input and drawn from a symmetric Bernoulli distribution, where each component takes the value $\pm 1$ with equal probability. In practice, the magnitude of the perturbation is matched to the dynamic range of the input image and subsequently scaled by a small constant $\beta$ to ensure that each step represents a local perturbation within the valid range expected by the model. We find that, as long as the perturbations remain symmetric about zero, the specific noise pattern or scale used does not significantly affect the final attribution or conclusions drawn, as discussed in Supplementary Material Section A3. The resulting stochastic gradient $\nabla f(\boldsymbol{\sigma}_k)$ is then used to update the input according to Eq. (3), driving the model's confidence in the predicted class toward zero. Analyses of the sensitivity to step-size ($\alpha$) and perturbation magnitude ($\beta$) parameters are provided in the Supplementary Material A4; in practice, we find that the method remains robust across a broad range of constant values ($0.01 \leq \beta \leq 0.5$, $\alpha \leq 1$). Iteratively applying this procedure produces a path through input space that follows the model's confidence landscape, forming the self-guided path. This path is illustrated in Figure 4, showing the original input, belonging to the 'Goldfish' class, being progressively perturbed toward a minimum in the confidence landscape of that same class. Unlike IG$^2$, which leverages

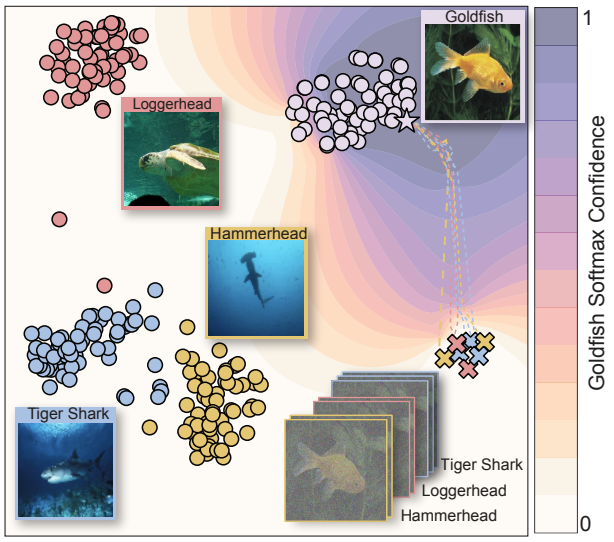


Figure 4. t-SNE plot of the SIGMA path. The image to be explained belongs to the Goldfish class and SIGMA efficiently minimises the model's confidence in the Goldfish class. Seven SIGMA paths are shown with dashed lines with the crosses indicating the end points of the path in feature space and resulting images of minimised confidence in Goldfish shown. The model now considers these images of belonging to different classes (Loggerhead, Tiger Shark and Hammerhead) without the images being near those classes in the feature space.

a representation layer to guide the input toward a different class in feature space, SIGMA traces a locally minimal path within the model's confidence landscape, connecting the input to a nearby low-confidence region. These final inputs are misclassified due to the collapse of class confidence, yet they remain outside their newly predicted class clusters in feature space.

### 3.2. Attribution and Weighting

After each update of $\boldsymbol{\sigma}_k$, the perturbed input is passed through the model, and a pixel-wise gradient, $g_k^{(i)} = \frac{\partial f(\boldsymbol{\sigma}_k)}{\partial \sigma_k^{(i)}}$ is calculated with respect to the model output for the target class, consistent with other gradient-based methods. We refer to the gradients across all pixels as the gradient map, $\boldsymbol{g}_k$, depicted in Figure 1. These gradient maps are then weighted such that the sum of the pixel gradients for that step is equal to the model's drop in confidence between iterations as motivated in Section 1. These weighted gradient maps are then summed to form the final attribution map. We formalise this accumulation as follows:

**Definition 1** (SIGMA Attribution)**.** *Let $N$ denote the total number of iterations, $\boldsymbol{\sigma}_k$ the input at the current iteration on the SIGMA path, $f$ the model's prediction function for the target class and $\boldsymbol{g}_k$ the gradient map at iteration k. The attribution map at iteration k is therefore:*

$$\boldsymbol{\phi}^{SIGMA} = \sum_{k=1}^{N} \boldsymbol{g}_k \times \frac{f(\boldsymbol{\sigma}_{k-1}) - f(\boldsymbol{\sigma}_k)}{\sum_{i'} g_k^{(i')}}. \tag{4}$$

*Here we use $i'$ to index over all of the pixel-wise gradients in the gradient map at iteration $k$.*

Pseudocode outlining the full implementation of the SIGMA path creation and resulting attribution is available in Section A1 of Supplementary Material.

### 3.3. Axiomatic Properties

Path methods are known to satisfy a number of desirable axioms, formally introduced in [25], as a substitute for concrete ground-truth comparisons. As discussed, the formulation of SIGMA ensures gradients are directly scaled by the model output at each stage of the path. This also ensures the axiom of Completeness, that summing over all pixels in the attribution map equals the total change in model confidence along the path, is inherently satisfied. Summing the attributions of each pixel $\phi_i^{\text{SIGMA}}$ Eq.4 becomes:

$$\sum_i \phi_i^{\text{SIGMA}} = \sum_{k=1}^{N} \frac{f(\boldsymbol{\sigma}_{k-1}) - f(\boldsymbol{\sigma}_k)}{\sum_{i'} g_k^{(i')}} \sum_i g_k^{(i)}. \tag{5}$$

Setting the index of each pixel in the gradient map $i'$ to the index of each pixel in the image $i$, the sum over all pixels in the final attribution map equal to the total drop in confidence between the first ($k = 0$) and final ($k = N$) iterations. Further discussion and proofs of the other axioms SIGMA satisfies, including Sensitivity (a), Sensitivity (b), Implementation Invariance and Symmetry can be found in Section A2 of Supplementary Material.

## 4. Experimental Evaluation

### 4.1. Datasets

SIGMA is validated across three diverse computer vision datasets: ImageNet—a widely used benchmark dataset containing 1,000 object classes [38]; the Disease Risk Estimation Dataset (Disease-XAI); and the Security Check Dataset (Security-XAI), both of which are part of Saliency-Bench [39]. Saliency Bench is a set of benchmark datasets created to test explainable AI models. They are accompanied by human annotations, enabling visual comparisons in lieu of ground truth attributions and allow for quantitative alignment metrics for a more robust comparison. The Disease-XAI dataset comprises 5250 annotated CT scans from lung cancer screenings, derived from LIDC-IDRI [40], with radiologist agreement used to define positive nodule regions. The CNN classifier, MobileNetV2, was fine-tuned on Disease-XAI images to distinguish nodule and non-nodule cases and attributions were evaluated against the annotated lesion masks. The Security-XAI dataset contains

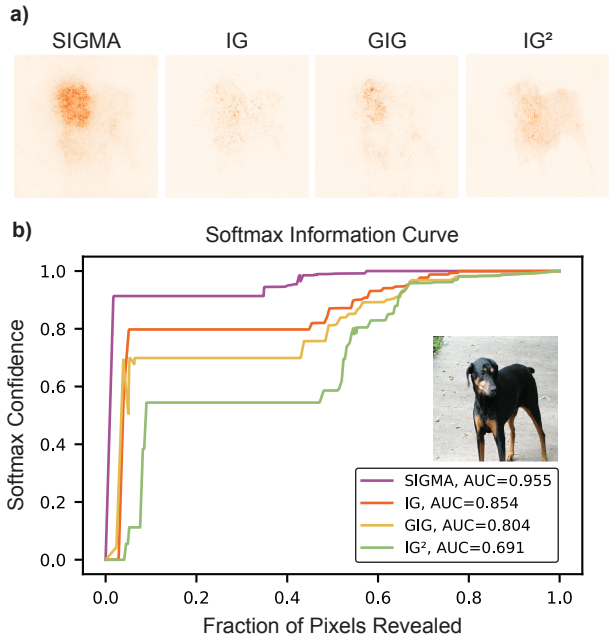


Figure 5. Comparison of attribution quality using the Softmax Information Curve (SIC) and visual attributions for InceptionV3 predicting the class 'Dobermann'. (a) Qualitative maps show that SIGMA produces sharper, more spatially focused attributions. (b) SIGMA achieves the highest SIC AUC, indicating stronger alignment between revealed pixels and model confidence.

17654 X-ray baggage images from the Sixray dataset [41], annotated by security experts to identify prohibited items. An InceptionV3 model was fine-tuned for classification of the five prohibited items present in the dataset.

### 4.2. Methods

We compare SIGMA with three established path-based attribution methods: the straight-line path of IG [25], the adaptive path of GIG [30], and the counterfactual formulation of IG$^2$ [31]. Due to its stochastic nature, the exact SIGMA trajectory from the original input to a low-confidence variant varies across runs. Averaging over multiple paths yields smoother and more stable attributions with a trade off with computational time. These experiments choose $n = 7$ paths as an appropriate trade off between attribution quality and computational time. (see Supplementary Section A6 for full discussion). Unlike IG and GIG, which require a fixed number of integration steps (200 in this study, following [25]), SIGMA terminates dynamically once model confidence in the target class falls below 1%. Step size and perturbation magnitude parameters, $\alpha$ and $\beta$ were set to 0.1 and 0.5 respectively for all experiments, further explanation on this parameter selection is available in Supplementary Section A4. For IG and GIG, a black image was used for the baseline as recommended in prior

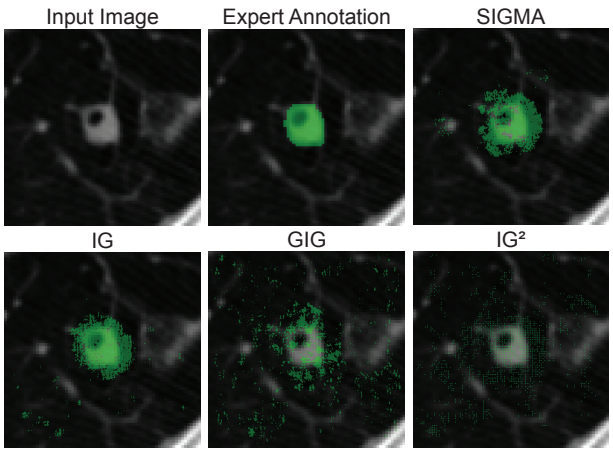


Figure 6. Binary feature attributions overlaid on a positive nodule image from the Disease-XAI dataset. The annotated region (green) indicates the nodule area agreed upon by at least 50% of radiologists. Attribution maps are shown in green to enable direct visual comparison with the annotation. Both IG and SIGMA produce sharp, spatially aligned attributions consistent with the annotated region, whereas Guided IG and IG$^2$ yield sparser, less localised responses.

work [25, 30]. IG$^2$ requires both a counterfactual reference and a representation layer; we use random counterfactuals and the penultimate layer, following the setup in [31].

### 4.3. Qualitative Comparison

Qualitative comparisons between SIGMA and the gradient-based methods outlined above are visualised for one example from ImageNet in Figure 5(a), Disease-XAI in Figure 6 and Security-XAI in Figure 7. Across datasets, SIGMA produces attributions that are spatially coherent, less noisy, and closely aligned with annotated regions. In this work we also extend our comparison to Vision Transformers (ViTs), see Figure 8, to assess the applicability of path-based attribution methods beyond convolutional networks and highlight the potential for attributions at a finer resolution than the transformer's own patch-limited attention maps, for this experiment, the ViT's own attention map was computed via attention rollout [42], aggregating multi-head attention across layers. Note that IG$^2$ was omitted from this study as it did not yield meaningful attributions for Vision Transformers. This may be related to architectural differences between convolutional and transformer-based models [32] as IG$^2$ requires a representation layer.

### 4.4. Quantitative Evaluation

We evaluate SIGMA using the SIC-AUC metric discussed in Section 1 and two techniques outlined in Saliency-Bench [39]. The analysis techniques in Saliency-Bench are designed to assess alignment, that is, how closely attribution

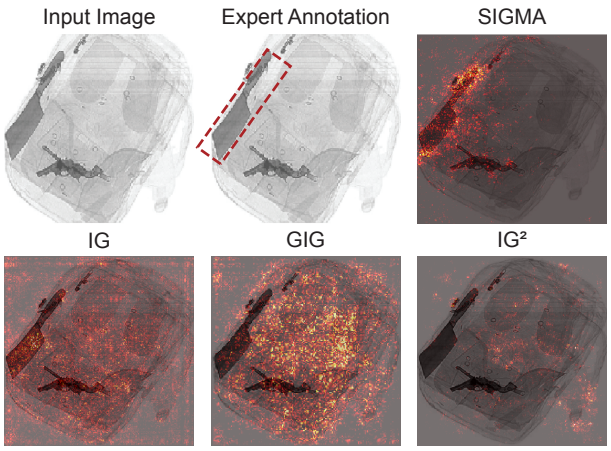


Figure 7. Comparison of attribution maps on an annotated X-ray baggage scans from the Security-XAI dataset with attributions overlayed for the prediction of the class 'Knife'. SIGMA highlights the annotated region with higher spatial precision and less background noise than IG, GIG, and $IG^2$.

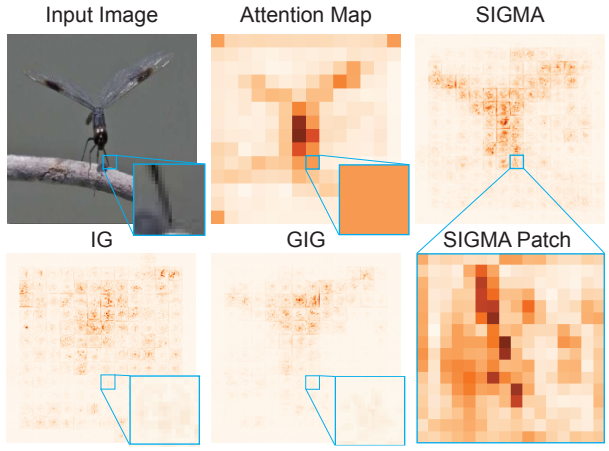


Figure 8. Attribution comparison for the ViT-B/16 Vision Transformer architecture, introduced in [43] with the input predicting the class 'Dragonfly'. SIGMA aligns with the model's attention map while also revealing sub-patch structure. As shown with one patch enlarged and compared for each method.

maps correspond to human understanding. Following the evaluation framework, we report two metrics: mean Intersection over Union (mIoU), which measures spatial overlap between attribution maps and human annotations; and the Pointing Game (PG), which checks whether the maximum attribution falls within the annotated region. For mIoU, attribution maps are thresholded to produce binary masks, where values closer to one indicate greater overlap. For PG, values closer to one imply that a higher percentage of attributions have their maximum value within the annotated region. Further discussion of these metrics, along with their formal definitions, is provided in the Section A7 of the Supplementary Material.

Across all evaluated datasets and network architectures, SIGMA achieves comparative scores to existing path-based attribution methods (Tables 1–2). On the ImageNet benchmarks, SIGMA consistently attains high SIC-AUC values, indicating strong faithfulness between attribution and model confidence. Similarly, on the XAI-Bench datasets, SIGMA shows consistent scores across the two networks and real-world datasets. These results suggest that SIGMA generalises effectively across both CNN and ViT architectures, and to more complex real-world datasets providing reliable explanations without the need for a baseline or counterfactual reference.

Table 1. Comparing the average SIC-AUC scores for attributions of 1000 ImageNet images tested across three CNNs and one Vision Transformer. Highest score is highlighted in bold.

| Method | ResNet50 | InceptionV3 | MobileNetV2 | ViT |
|---|---|---|---|---|
| IG | 0.53 | 0.56 | 0.55 | **0.76** |
| GIG | 0.58 | 0.53 | 0.53 | 0.73 |
| $IG^2$ | 0.59 | 0.54 | 0.48 | - |
| SIGMA | **0.60** | **0.60** | **0.56** | 0.75 |

Table 2. Quantitative comparison of attribution methods on XAI-Bench datasets. 1000 images were taken from each dataset, with the average metric across all attributions reported below. The highest scores for each metric are highlighted in bold.

| Dataset | Network | Method | mIOU | PG | SIC |
|---|---|---|---|---|---|
| **Nodule-XAI** | MobileNetV2 | IG | **0.23** | **0.94** | 0.72 |
| | | GIG | 0.17 | 0.80 | 0.74 |
| | | $IG^2$ | 0.02 | 0.34 | 0.50 |
| | | SIGMA | 0.20 | 0.58 | **0.76** |
| **Security-XAI** | InceptionV3 | IG | 0.07 | 0.23 | 0.60 |
| | | GIG | 0.12 | 0.53 | 0.60 |
| | | $IG^2$ | 0.05 | 0.14 | 0.52 |
| | | SIGMA | **0.15** | **0.58** | **0.62** |

## 5. Confidence Bounds

Employing multiple stochastic paths enables estimation of the standard error of the attribution map, providing a quantitative measure of uncertainty in the explanation. The standard error reflects the consistency of the attributions across paths and is computed as the standard deviation of attributions divided by the square root of the number of paths, $\Delta\phi/\sqrt{n}$. As shown in Figure 9, this measure highlights regions where the model's explanations are most stable and reveals areas of higher variability where the attribution is less certain. Increasing the number of paths reduces this variance, yielding smoother and more coherent attribution

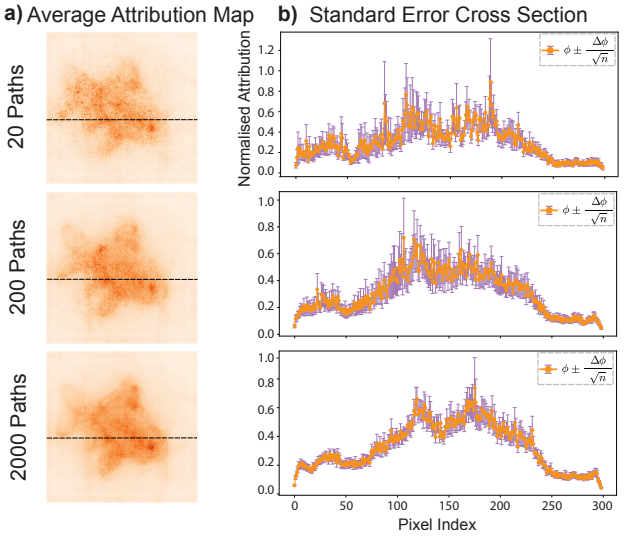


Figure 9. (a) Average attribution maps for 20, 200, and 2000 paths for the 'Goldfish' example from Fig. 1. The attribution maps become cleaner as the number of paths is increased, at the cost of computational time. The dotted lines indicate the cross sections shown in (b). The standard error (purple) is given by $\Delta\phi/\sqrt{n}$, where $n$ is the number of paths.

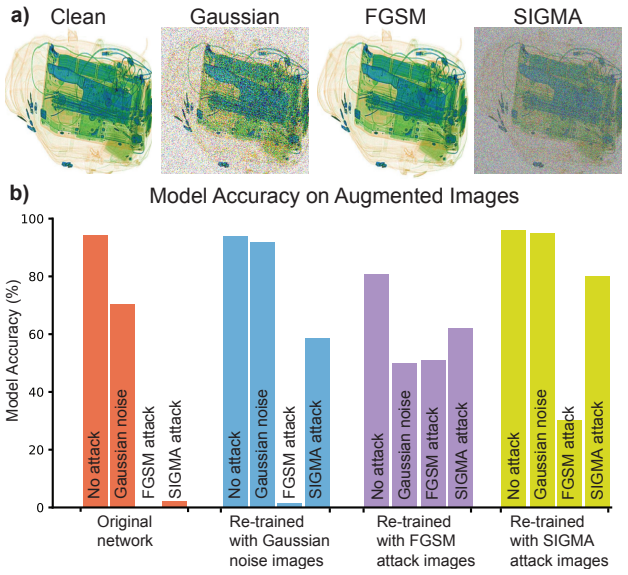


Figure 10. (a) Example augmentations applied to the Security-XAI test dataset, including Gaussian noise, FGSM perturbations, and SIGMA-generated zero-confidence inputs. (b) Model accuracy on test images augmented with each method after retraining. Retraining with SIGMA augmentations (green) improves robustness to both Gaussian and adversarial noise while maintaining high clean-image accuracy.

maps that capture a fuller picture of the model's decision process. Further, the relative standard error could be used as a stopping criterion to determine the number of SIGMA paths, such that all attribution maps are standardised in terms of their confidence. See Supplementary Material Section A5 for further discussion on path averaging and confidence bounds.

## 6. Increasing Model Robustness

As discussed in Section 1, the SIGMA path naturally generates zero-confidence variants of the input that resemble noisy versions of the original image. These images remain recognisable to humans but receive a near-zero confidence from the network, mimicking an adversarial attack. Here, we evaluate the potential of these images used as data augmentations to improve model robustness to both noise and adversarial attacks. A total of 500 training images from the Security-XAI dataset were augmented under three conditions; Gaussian noise, a Fast Gradient Sign Method (FGSM) adversarial attack [44], and a SIGMA attack, these were then each used to retrain the InceptionV3 model used in Section 4, creating three augmented models.

Each retrained model was then evaluated on 100 images drawn from the test dataset under four conditions, the clean images with no augmentation applied, images with Gaussian noise applied, FGSM attacked images and SIGMA attacked images. This enables a direct comparison of robustness across different retraining strategies and test conditions. Figure 10 shows that the model retrained with SIGMA augmentations achieves the highest combined accuracy across all perturbation types. It maintains clean-image performance, while also providing robustness to Gaussian noise and FGSM attacks, suggesting that SIGMA provides a balanced regularisation between noise-based and adversarial strategies.

## 7. Conclusion

This work introduced SIGMA, a baseline-free, path-based attribution method that explores the model's confidence landscape through stochastic perturbations. By dynamically tracing confidence descent paths without predefined baselines or counterfactuals, SIGMA produces interpretable and faithful attribution maps across diverse architectures and datasets. The method satisfies key axiomatic properties and enables path averaging and uncertainty estimation, providing a quantitative measure of explanation reliability. Beyond interpretability, the same mechanism generates low-confidence variants of inputs, offering a general framework that has the potential to increase network robustness to both noise and targeted attacks.

## References


[1] Holger A Haenssle, Christine Fink, Roland Schneiderbauer, Ferdinand Toberer, Timo Buhl, Andreas Blum, Aadi Kalloo, A Ben Hadj Hassen, Luc Thomas, and Alexander Enk. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of oncology*, 29(8):1836–1842, 2018. 1

[2] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, and Ara Darzi. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[3] Miquel Alfaras, Miguel C Soriano, and Silvia Ortín. A fast machine learning model for ecg-based heartbeat classification and arrhythmia detection. *Frontiers in Physics*, 7:103, 2019.

[4] Hyo-Eun Kim, Hak Hee Kim, Boo-Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun-Kyung Kim. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health*, 2(3):e138–e148, 2020. 1

[5] Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. Deep learning for financial applications: A survey. *Applied soft computing*, 93:106384, 2020. 1

[6] Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance*, 34(11):2767–2787, 2010.

[7] Jin Xiao, Yu Zhong, Yanlin Jia, Yadong Wang, Ruoyi Li, Xiaoyi Jiang, and Shouyang Wang. A novel deep ensemble model for imbalanced credit scoring in internet finance. *International Journal of Forecasting*, 40(1):348–372, 2024. 1

[8] Samet Akcay, Mikolaj E Kundegorski, Chris G Willcocks, and Toby P Breckon. Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. *IEEE transactions on information forensics and security*, 13(9):2203–2215, 2018. 1

[9] Xiang Bai, Mingkun Yang, Tengteng Huang, Zhiyong Dou, Rui Yu, and Yongchao Xu. Deep-person: Learning discriminative deep features for person re-identification. *Pattern Recognition*, 98:107036, 2020.

[10] Stirling Scholes, Alice Ruget, Feng Zhu, and Jonathan Leach. Human pose inference using an elevated mmwave fmcw radar. *IEEE Access*, 2024. 1

[11] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016. 1

[12] Zhiyu Huang, Chen Lv, Yang Xing, and Jingda Wu. Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. *IEEE Sensors Journal*, 21(10):11781–11790, 2020.

[13] Xuerui Dai, Xue Yuan, and Xueye Wei. Tirnet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence*, 51(3):1244–1261, 2021.

[14] Stirling Scholes, Alice Ruget, Germán Mora-Martín, Feng Zhu, Istvan Gyongy, and Jonathan Leach. Dronesense: The identification, segmentation, and orientation detection of drones via neural networks. *IEEE Access*, 10:38154–38164, 2022. 1

[15] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017. 1

[16] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120, 2019. 1

[17] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, and Richard Benjamins. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020. 1

[18] Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 652–663, 2021. 1

[19] MD Zeiler. Visualizing and understanding convolutional networks. In *European conference on computer vision/arXiv*, volume 1311, 2014. 1

[20] V Petsiuk. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 1

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016. 1

[22] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 1

[23] Ludwig Schallner, Johannes Rabold, Oliver Scholz, and Ute Schmid. Effect of superpixel aggregation on explanations in lime–a case study with biological data. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 147–158. Springer, 2019. 1

[24] Y Wang, T Zhang, X Guo, and Z Shen. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415*, 2024. 1

[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1, 3, 5, 6, 4

[26] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020. 1, 3

[27] Vivek Miglani, Narine Kokhlikyan, Bilal Alsallakh, Miguel Martin, and Orion Reblitz-Richardson. Investigating sat-

uration effects in integrated gradients. *arXiv preprint arXiv:2010.12697*, 2020. 1, 3

[28] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017. 1

[29] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020. 1

[30] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5050–5058, 2021. 1, 3, 6, 2

[31] Yue Zhuo and Zhiqiang Ge. IG2: Integrated Gradient on Iterative Gradient Path for Feature Attribution . *IEEE Transactions on Pattern Analysis Machine Intelligence*, 46(11):7173–7190, 2024. 1, 2, 6

[32] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021. 2, 6

[33] James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 1992. 2, 4

[34] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4948–4957, 2019. 3

[35] Huazheng Du, Guoye Chen, Xuegang Hu, Na Xia, and Biaodian Xu. Simultaneous perturbation stochastic approximation–based radio occultation data assimilation for sensing atmospheric parameters. *International Journal of Distributed Sensor Networks*, 14(12):1550147718815848, 2018. 4

[36] Christopher Ferrie. Self-guided quantum tomography. *Physical review letters*, 113(19):190404, 2014.

[37] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017. 4

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

[39] Yifei Zhang, Siyi Gu, James Song, Bo Pan, and Liang Zhao. Xai benchmark for visual explanation, 2023. 5, 6, 4

[40] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. 5

[41] Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao, and Qixiang Ye. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2119–2128, 2019. 6

[42] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 6

[43] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[44] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 8

[45] Eric J. Friedman. Paths and consistency in additive cost sharing. *International Journal of Game Theory*, 32(4):501–518, 2004. 1