Manifold Integrated Gradients: Riemannian Geometry for Feature Attribution

Eslam Zaher [1 2] **Maciej Trzaskowski** [1 3] **Quan Nguyen** [1 3 4] **Fred Roosta** [1 2]

Abstract

In this paper, we dive into the reliability concerns of Integrated Gradients (IG), a prevalent feature attribution method for black-box deep learning models. We particularly address two predominant challenges associated with IG: the generation of noisy feature visualizations for vision models and the vulnerability to adversarial attributional attacks. Our approach involves an adaptation of path-based feature attribution, aligning the path of attribution more closely to the intrinsic geometry of the data manifold. Our experiments utilise deep generative models applied to several real-world image datasets. They demonstrate that IG along the geodesics conforms to the curved geometry of the Riemannian data manifold, generating more perceptually intuitive explanations and, subsequently, substantially increasing robustness to targeted attributional attacks.

1. Introduction

As complexity of deep learning models continues to accelerate, ensuring their trustworthiness becomes paramount. Explainability has emerged as an important research field, making the behaviour of black-box deep learning more transparent. In this context, feature attribution methods (Sundararajan et al., 2017; Springenberg et al., 2015; Ribeiro et al., 2016; Lundberg & Lee, 2017; Simonyan et al., 2014; Bach et al., 2015) represent a family of explainability techniques designed to attribute a model's prediction to the most salient features in the input. These methods allow for the assessment of whether a model's decision is based on quantifiable and valid reasons, specifically in terms of the most influential factors that contribute to that decision.

[1]ARC Training Centre for Information Resilience (CIRES), Brisbane, Australia [2]School of Mathematics and Physics, University of Queensland, Brisbane, Australia. [3]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia [4]QIMR Berghofer Medical Research Institute, Brisbane, Australia. Correspondence to: Eslam Zaher <e.zaher@uq.edu.au>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Gradient-based feature attribution methods (Simonyan et al., 2014; Springenberg et al., 2015; Bach et al., 2015) are commonly used, as they are more computationally efficient and faithful to the model's internals, as opposed to other perturbation-based techniques (Ribeiro et al., 2016; Lundberg & Lee, 2017). However, these methods suffer from two main drawbacks. First, the saliency maps produced for vision models often exhibit perceptual noise. Second, they are susceptible to adversarial attributional attacks.

Integrated gradients (IG) and other path-based methods are widely adopted because they satisfy axiomatic properties that are desirable in feature attribution methods. However, the choice of the path of attribution, and/or the baseline, impacts the quality and robustness of the generated explanations. For vision models, IG has faced specific criticism for accumulating noise along the integration path, diminishing the human-perceptual quality of the resulting explanations. A few studies attributed the main source of this noise to the ad hoc choice of the linear path of attribution. Kapishnikov et al. (2021) indicated that the linear path in IG is agnostic to the model output surface, leading to the accumulation of high-norm gradients assigned to irrelevant image pixels. Miglani et al. (2020) showed that gradients in saturated regions along the integration path, where the model output plateaus, can lead to disproportionate attributions. Yang et al. (2023) analyzed gradients at each point on the path, decomposing them into relevant and noise directions. Their results showed that including the noise direction in the path integral contributes considerably to the overall noise in the explanations.

Several methods have been proposed to mitigate the noise in the IG saliency maps. Kapishnikov et al. (2021) greedily optimized the path of attribution by guiding it through the model output surface. However, the resulting path from this method may stray into regions associated with adversarial examples due to its significant deviation from the straight-line path. Rather than utilizing the linear path in the image space, Xu et al. (2020) integrated gradients along a continuum of blurred images. While this approach helps in reducing noise, it can also lead to a loss of fine-grained details due to the blurring effect. Split IG (Miglani et al., 2020) excludes noisy saturated regions; however, it breaks the axioms fulfilled by IG. Other methods attempt to improve attributions by manipulating the input image and/or

the baseline. SmoothGrad (Smilkov et al., 2017) averages attribution maps of noisy instances of the input image to generate less noisy feature visualisations. Alternative baselines (Sturmfels et al., 2020; Fong & Vedaldi, 2017; Smilkov et al., 2017) or even distributions of baselines (Lundstrom et al., 2022; Erion et al., 2021) have also been explored.

While the aforementioned methods aim at mitigating the noise issue in IG, there has been comparatively less emphasis on addressing IG's susceptibility to adversarial attributional attacks. Ghorbani et al. (2019) demonstrated that minimal perturbations to the input image can generate unstructured attributions due to the rapidly changing gradients at the decision boundary. IG has been shown to suffer from targeted attributional attacks, which can manipulate images to generate saliency maps that mimic explanations of arbitrary images, while maintaining a constant model output. Vulnerability to attributional attacks even apply to local model-agnostic methods (Slack et al., 2020) and global explanations (Heo et al., 2019; Baniecki et al., 2023; Laberge et al., 2022). As a remedy to these attributional attacks, common approaches employ either one of two main strategies: (1) informing the classifiers about adversarial examples via adversarial training (Madry et al., 2018; Chalasani et al., 2020), (2) modifying the objective function to maximise correlation of feature attributions between original and perturbed inputs (Dombrowski et al., 2019; Chen et al., 2019). There are also hybrid approaches that combine elements of both strategies (Ivankay et al., 2022; Chen et al., 2019; Wang & Kong, 2022).

The bulk of work in adversarial robustness has substantiated that robust classifiers exhibit input-gradients that are semantically aligned with the human perception - a phenomena known as perceptually aligned gradients (PAGs) (Kaur et al., 2019; Ganz et al., 2023; Shah et al., 2021; Madry et al., 2018). Robustification using adversarial training is widely considered the most effective approach to rectify the model sensitivity to input perturbations, leading to both robust predictions and enhanced feature visualisations. The sharpness and robustness of feature visualizations from these classifiers is essentially a manifestation of PAGs. A few studies have shown that the data manifold is critical to achieving both robust classifiers and perceptual feature attributions. Srinivas et al. (2023) argued that for models to exhibit PAGs, they need to be more robust off the data manifold than on it. The resulting PAGs from off-manifold robustness are found to align closely with the data manifold. The interplay between the data manifold, PAGs, and gradient-based feature attribution methods has also been explored. Bordt et al. (2023) showed that the alignment of gradient-based saliency vectors with the data manifold corresponds to increased perceptual quality. Likewise, as models become more robust, their gradients tend to align more closely with the data manifold.

Contributions. Motivated by Srinivas et al. (2023); Bordt et al. (2023), we take a data-manifold guided approach to simultaneously address both drawbacks of IG, i.e., the perceptual noise and the vulnerability to targeted attributional attacks. The main contributions of our work are as follows:

(i) We introduce Manifold IG (MIG), a novel path-based attribution method that, by integrating along the geodesics of a latent Riemannian manifold, respects the curved geometry inherent to the underlying data manifold.

(ii) Using deep generative models on real-world image datasets, we show that not only does MIG yield perceptually aligned feature visualizations but it also makes feature attributions more robust to targeted attributional attacks.

2. Background

We now briefly review several path-based feature attribution methods and provide some essential background necessary for our presentation.

2.1. Path-based Feature Attribution Methods

For target image x, baseline x', and a classifier F, Gradient-based explainability methods are defined in terms of the derivative of F with respect to the input. More specifically, *path attribution methods* are a class of methods defined in terms of a path $\gamma(t) := \gamma(x, x', t)$ connecting a baseline x' and a target input x by a continuous, smooth curve. In this light, path attribution methods are also baseline methods, and can be defined as

$$\text{PathAttr}_j(x, x', \gamma) := \int_0^1 \frac{\partial F(\gamma(t))}{\partial \gamma_j(t)} \frac{\partial \gamma_j(t)}{\partial t}\, dt.$$

Depending on the choice of the baseline and the path of attribution, various approaches emerge. We now outline a few of these methods.

Integrated Gradients (IG). the original path-based feature attribution method (Sundararajan et al., 2017) with the simplest linear path function $\gamma(t) = x' + t(x - x')$ as

$$\text{IG}_j(x, x') := (x_j - x'_j) \times \int_{t=0}^1 \frac{\partial F(x' + t(x - x'))}{\partial x_j} dt.$$

The baseline is commonly chosen to be either a black or white image in the case of vision models. IG aggregates contributions of features along a path from a baseline with an absolute absence of influential features, and progressively increases the presence of the feature signal along the path until reaching full abundance at the target feature value.

Guided Integrated Gradients (GIG). Kapishnikov et al. (2021) showed that IG generates noisy attributions due to accumulation of noise along the linear path. They generalized the path function to be guided by the model output

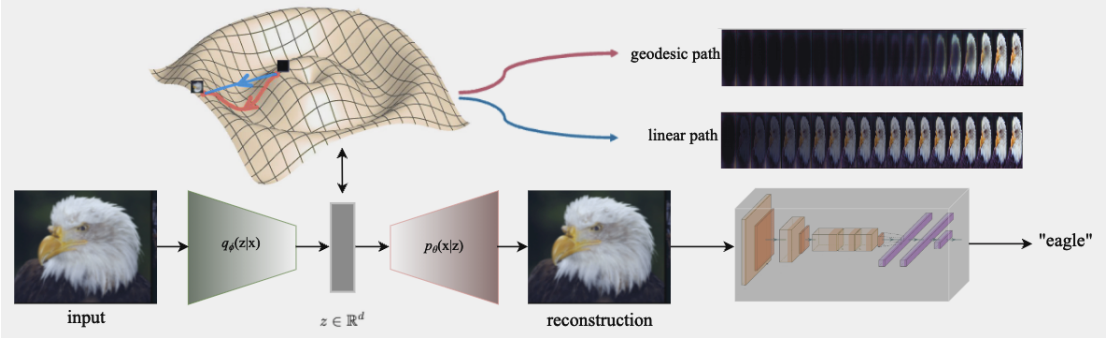

Figure 1. Schematic of our Setup: The underlying image data manifold is learned using a convolutional VAE. The latent space corresponds to a Riemannian manifold where the geodesic path (shown in red) between two points represents the shortest path in such curved geometry. The linear path (shown in blue) doesn't conform to the intrinsic geometry of the manifold and deviates into regions out of the manifold. Reconstructions from the VAE along with the labels are used to train a classifier, and the geodesic path is used as the path of attribution in our MIG as opposed to the linear path in the image space used in IG.

surface, allowing the path of attribution to avoid regions of saturated and high gradients. GIG can be expressed as

$$\text{GIG}_j(x, x') := \int_{t=0}^{1} \frac{\partial F(\gamma^F(t))}{\partial \gamma_j^F(t)} \frac{\partial \gamma_j^F(t)}{\partial t}\, dt,$$

where $\gamma^F(t) := \gamma(x, x', F, t)$ represents the path guided by the model, which is greedily optimized to minimize the impact of high-norm model gradients.

Blur IG. This method can be viewed as a variation of IG on a path that transitions from a fully blurred image to the original image (Xu et al., 2020), and is given by

$$\text{BlurIG}(m, n) = \int_{t=\infty}^{0} \frac{\partial F(L(m, n, t))}{\partial L(m, n, t)} \frac{\partial L(m, n, t)}{\partial t} dt,$$

where m, n are the pixel indices in the image, and L is a 2D Gaussian blur kernel with variance t as

$$L(m, n, t) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \frac{1}{\pi t} e^{-\frac{m^2+n^2}{t^2}} x(m-k, n-l).$$

2.2. Attributional Attacks

The susceptibility of feature attribution methods to adversarial attributional attacks raises concerns about their reliability. Ghorbani et al. (2019) demonstrated that adding systematic human-imperceptible perturbations to the input image can cause amorphous change in the feature maps, while maintaining the same output class. Dombrowski et al. (2019) proposed a targeted attack to generate feature explanations that can match any arbitrary target feature map. They showed that the model output surface in ReLU-based image classifiers exhibits high curvature, causing increased vulnerability of the gradient-based explanations. To enhance robustness of explanations, they smoothed the model's output curvature by replacing the ReLU activations in the model with SoftPlus at the time of generating attributions.

Among the common attacks in this context are top-k and targeted attributional attacks.

Top-k Attributional Attack. This approach generates an explanation that reduces the attribution score of the highest k features in the original feature map. Ghorbani et al. (2019) defines the top-k attack as $x^*_{\text{adv}} = \arg\min_{x_{\text{adv}}} C(I(x), I(x_{\text{adv}}))$ subject to the constraints that $\|x_{\text{adv}} - x\|_\infty \leq \epsilon$ and $F(x_{\text{adv}}) = F(x)$, where I is the attribution method of interest, $C(x, x_{\text{adv}}) = \sum_{j \in K} I_j(x)$, K is the set of top k features in x_{adv}, and $\|.\|_\infty$ denotes the infinity norm on vectors.

Targeted Attributional Attack. This method manipulates the input image to generate attributions that resemble feature maps corresponding to a different target image (Dombrowski et al., 2019). It can be formulated as $x^*_{\text{adv}} = \arg\min_{x_{\text{adv}}} \mathcal{L}(x_{\text{adv}})$ where $\mathcal{L}(x_{\text{adv}}) = \|I(x_{\text{adv}}) - I(x_{\text{target}})\|_2^2 + \gamma \|F(x_{\text{adv}}) - F(x)\|_2^2$, with γ and x_{target} being, respectively, a hyperparameter that controls the relative weight of the two summands and the target image. Here, $\|.\|_2$ signifies the Euclidean norm.

2.3. VAEs for Generative Manifold Learning

A substantial body of literature on manifold learning relies on neural autoencoders (AEs) (Hinton & Zemel, 1993; Vincent et al., 2008). The underlying hypothesis guiding the success of manifold learning is that even though the data may exist in a high-dimensional space, it effectively lies on or near an embedded low-dimensional manifold (Fefferman et al., 2016; Bengio et al., 2013; Brahma et al., 2016).

Unlike most traditional AEs that lack structure over the

latent space, variational autoencoders (VAEs) are the probabilistic variants with a regularised, generative latent structure (Kingma & Welling, 2022; Higgins et al., 2016; Burgess et al., 2018). VAEs aim not only to learn a compressed representation but also to model the underlying probability distribution of the data in a lower-dimensional latent space. This latent space is typically assumed to adhere to a prior distribution, often modeled as a multivariate Gaussian. This allows VAEs to capture the intrinsic geometrical and statistical structure of the data using more informative latent representations. The learned latent space is then used to generate new data points similar to the training data, making VAEs suitable for tasks such as image generation and reconstruction, interpolation, and out of distribution detection (Ran et al., 2022; Cristovao et al., 2020).

High-dimensional data such as real images are assumed to have a non-Euclidean latent structure that constitutes the natural image manifold (Zhu et al., 2016). Nevertheless, the data is often perceived through the lens of Euclidean geometry as it offers definite inner products and explicit distance metrics, making manipulation of data more accessible. Shao et al. (2018) showed that nonlinear VAEs induce a Riemannian metric over the latent space. Similarly, Arvanitidis et al. (2018) demonstrated that the induced Riemannian metric leads to faithful latent-space statistical estimates, smooth interpolations, and better generalisations. This motivated a compilation of methods to learn non-Euclidean latent structures, e.g., hyperspherical (Davidson et al., 2018), hyperbolic (Tifrea et al., 2018; Mathieu et al., 2019; Cho et al., 2023), mixed-curvature (Skopek et al., 2019) and Riemannian weighted submanifolds (Miolane & Holmes, 2020).

3. Riemannian Geometry for Feature Attribution

We now present our proposed feature attribution method. To that end, we present elements of differential geometry as they apply to deep generative models. We demonstrate how VAEs induce a Riemannian metric over the latent space. Utilizing the induced metric, we use an algorithm to compute geodesic paths that conform to the curved structure of the data manifold. Following this, we introduce Manifold Integrated Gradients (MIG) for feature attribution by integrating model gradients along these geodesics.

3.1. Latent Space Geometry in Deep Generative Models

A smooth manifold $\mathcal{M}$ is a topological manifold with a smooth structure that is locally homeomorphic to Euclidean space (Lee, 2014). This implies that around any point $z \in \mathcal{M}$ on the manifold, $\mathcal{M}$ resembles $\mathbb{R}^d$, enabling to extend differential calculus on the manifold. On smooth manifolds, concepts like length, inner products, and shortest paths, known as geodesics, take on a more complex nature compared to their Euclidean counterparts.

The governing mathematical apparatus that generalizes these concept to manifolds and gives a formal framework to compute them is the *Riemannian metric* (Lee, 2018). Recall that a tangent space of $\mathcal{M}$ at z, denoted by $T_z\mathcal{M}$, is a vector space spanning all the tangent vectors to z on $\mathcal{M}$. Since $T_z\mathcal{M}$ is a vector space, we can define an inner products on it. If this inner product varies smoothly with z, then it defines a Riemannian metric. More specifically, a Riemannian metric on a smooth manifold $\mathcal{M}$ assigns to each point $z \in \mathcal{M}$ an inner product $\langle\ ,\ \rangle_z : T_z\mathcal{M} \times T_z\mathcal{M} \to \mathbb{R}$ that varies smoothly on the manifold.A *Riemannian manifold* is a smooth manifold with a Riemannian metric (Lee, 2018).

Latent-variable generative models with smooth generator functions embody surface models (Arvanitidis et al., 2018). They incorporate a generative function: $g : \mathcal{M} \to \mathcal{X}$, which transforms a latent manifold $\mathcal{M}$ in the latent space $\mathcal{Z} \subset \mathbb{R}^d$ into a data manifold embedded in the data space $\mathcal{X} \subset \mathbb{R}^D$. In the VAE setting, the decoder acts as the generator function, and the latent space is chosen to have a lower dimensionality compared to the input space, i.e., $d \leq D$.

A smooth latent curve connecting points $z_0 \in \mathcal{M}$ and $z_1 \in \mathcal{M}$ can be parameterized by a function $\gamma : [0, 1] \to \mathcal{M}$ with $\gamma(0) = z_0$ and $\gamma(1) = z_1$. This is then mapped by the generative function to a corresponding smooth curve $g(\gamma) : [0, 1] \to \mathcal{X}$ on the data manifold. The length of the curve $g(\gamma)$ can be expressed as

$$\begin{aligned} L(g(\gamma)) &= \int_0^1 \left\| \frac{dg(\gamma(t))}{dt} \right\| dt \\ &= \int_0^1 \|J_g(\gamma(t))\gamma'(t)\|\, dt, \end{aligned}$$

where $J_g(\gamma(t)) = \partial g(z)/\partial z \mid_{z=\gamma(t)}$ is the Jacobian of g at $\gamma(t)$, and $\gamma'(t) = \partial\gamma(t)/\partial t$ is the velocity vector tangential to the latent curve. This can in turn be written as

$$\begin{aligned} L(g(\gamma)) &= \int_0^1 \sqrt{(J_g(\gamma(t)) \cdot \gamma'(t))^T \cdot (J_g(\gamma(t)) \cdot \gamma'(t))}\, dt \\ &= \int_0^1 \sqrt{\gamma'(t)^T G_g(\gamma(t))\gamma'(t)}\, dt, \end{aligned}$$

where $G_g(.) = J_g^T(.)J_g(.)$. Under certain architectural choices (Shao et al., 2018), $G_g(.)$ is symmetric positive definite and smooth on $\mathcal{M}$, and hence defines a Riemannian metric, which can in turn be used to calculate geodesics, i.e., the shortest path between two points on $\mathcal{M}$.

The length functional, $L(g(\gamma))$, is invariant under reparameterization (Jost, 2017, Lemma 1.4.3). Hence, to find curves of shortest length on $\mathcal{M}$, we can simply consider curves that are parameterized proportionally by arc length. Conse-

quently, it can be shown that the following energy functional

$$E(g(\gamma)) = \frac{1}{2}\int_0^1 \gamma'(t)^T G_g(\gamma(t))\gamma'(t)\,dt, \tag{1}$$

is essentially equivalent to $L^2(g(\gamma))$. Indeed, while in general, $L^2(g(\gamma)) \leq 2E(g(\gamma))$, the equality holds if and only if $\|dg(\gamma(t))/dt\|$ is constant (Jost, 2017, Lemma 1.4.2). This, gives rise to the following minimization problem

$$\min_{\gamma\in\Gamma} E(g(\gamma)), \tag{2a}$$

where

$$\Gamma = \Big\{\gamma : [0,1] \to \mathcal{M} \mid \gamma(0) = z_0,\ \gamma(1) = z_1,\ \text{ and } \|dg(\gamma(t))/dt\| \equiv \text{const}\Big\}, \tag{2b}$$

and whose solution γ^* is the shortest path, or geodesic, between z_0 and z_1 on $\mathcal{M}$. In this light, traversing along the shortest path from z_0 to z_1 on the latent manifold amounts to a smooth transition from $g(z_0)$ to $g(z_1)$ on the data manifold in the sense of (2); see Figure 2.

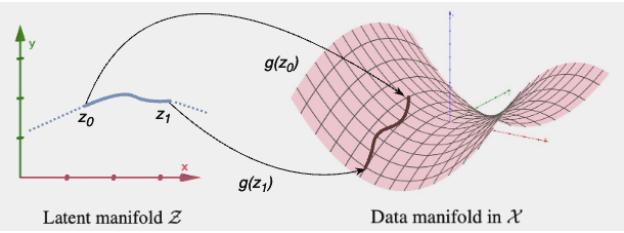

Figure 2. The surface model implied by the smooth generator function g mapping from the latent space $\mathcal{Z}$ to the data space $\mathcal{X}$. In this example, the latent manifold $\mathcal{M}$ is a one-dimensional embedded submanifold of $\mathbb{R}^2$ and the images lie on a two-dimensional embedded submanifold of $\mathbb{R}^3$. The geodesic on the latent manifold is mapped to a smooth curve on the data manifold, respecting the underlying geometry.

It can be shown that the curve $\gamma^*(t)$ is a solution to (2) if and only if it satisfies the system of second-order ordinary differential equations (Jost, 2017, Lemma 1.4.4)

$$\frac{d^2\gamma^k(t)}{dt^2} + \sum_{i,j}\Gamma_{ij}^k \frac{d\gamma^i(t)}{dt}\frac{d\gamma^j(t)}{dt} = 0, \tag{3}$$

where $k = 1,\dots,d$, and Γ_{ij}^k are the Christoffel symbols associated to the Riemannian metric tensor $G_g = (g_{ij})_{i,j=1,\dots,d}$. The Christoffel symbols are defined as

$$\Gamma_{ij}^k = \frac{1}{2}\sum_{\ell=1}^d g^{k\ell}\left(\frac{\partial g_{j\ell}}{\partial x^i} + \frac{\partial g_{i\ell}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^\ell}\right), \tag{4}$$

where $(g^{ij}) = (g_{ij})^{-1}$, i.e., $\sum_{\ell=1}^d g^{i\ell}g_{\ell j} = \delta_{ij}$.

Previous works have proposed different approaches to calculate geodesic paths for generative models. Shao et al. (2018) used a discretization of γ and a finite-difference formulation of the energy functional (1) to avoid the computational burden of the Euler-Lagrange system of equations (3). Arvanitidis et al. (2018) computed the Christoffel symbols (4) and then solved the second order system numerically to get geodesic paths for VAEs with stochastic generators. Arvanitidis et al. (2019) employed a scheme based on fixed-point iterations to solve the system in (3) without computing Jacobians that are usually ill-conditioned. In this work, we use a slight modification of Shao et al. (2018, Algorithm 1) to calculate geodesic paths; see Appendix A for details.

3.2. Integrated Gradients on the Data Manifold

Our generative-discriminative approach is depicted in Figure 1. We train a VAE to capture the underlying Riemannian data manifold for real image datasets. We feed-forward the whole datasets through the VAE and use the resulting reconstructions from the learned image data manifold along with the ground-truth labels to train a deep convolutional classifier. We then exploit the intrinsic geometry of the manifold to build a path-based feature attribution method by integrating gradients of the classifier's output along geodesics on the manifold. This leads to defining MIG along the geodesic path, γ^* from (2), for the j^{th} feature as

$$\text{MIG}_j(x,\gamma^*) := \int_0^1 \frac{\partial F(g(\gamma^*(t)))}{\partial g_j(\gamma^*(t))}\frac{\partial g_j(\gamma^*(t))}{\partial t}\,dt.$$

MIG employs the smoothest path between a baseline and a target in the sense of (2). In essence, the smooth generator maps the geodesic on the latent space to a highly smooth path on the data manifold that respects the underlying curved geometry. Further, like any path-based attribution method, MIG satisfies the axioms of completeness, sensitivity, and implementation invariance (Sundararajan et al., 2017).

4. Experiments

We now evaluate MIG in several contexts. In Section 4.1, we first demonstrate the effect of smooth interpolating paths adherent to the underlying geometry that underpins MIG. In Section 4.2, we compare various methods with MIG in terms of the perceptual alignment of their feature attribution maps. Finally, in Section 4.3, we investigate MIG in the context of targeted attributional attacks. The code for these experiments is available here.

Setup. For learning the image data manifold, we employ a convolutional VAE based on residual connections; see Appendix B for details. We extend the typical Evidence Lower Bound (ELBO) loss by adding a perceptual loss term (Hou

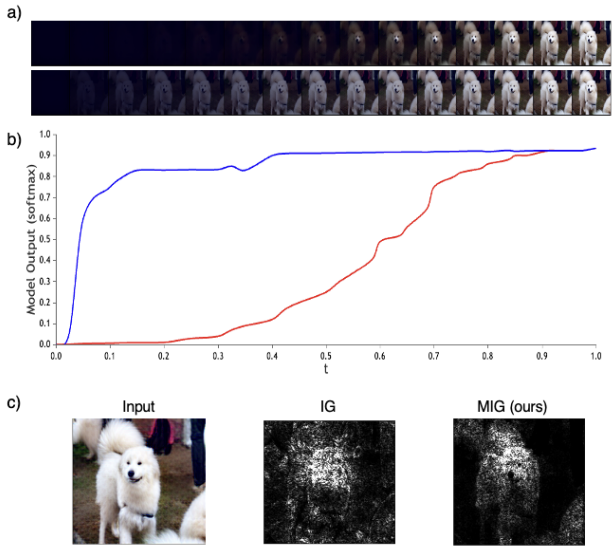

Figure 3. Mapped geodesic interpolation in MIG vs. linear interpolation in IG. (a) contrasts the smoothness of MIG's smooth path interpolants against IG's linear path from a black baseline. (b) displays a classifier response curves for each image on the paths, with MIG's smooth path (red) having a gradual response as key features show later on the path and IG's linear path (blue) showing rapid escalation, with a wide saturation region. (c) shows the corresponding feature visualizations. MIG produces more perceptually aligned and less noisy feature visualizations compared to IG.

et al., 2016) that emphasises feature-wise accuracy. Training VAEs with perceptual loss helps to mitigate the inherent blurriness in reconstructions as it preserves structural and perceptual details. We train the VAEs for 150 epochs with Adam (Kingma & Ba, 2014). We then use the image reconstructions along with the labels to train classifiers based on pretrained models.

For showing the perceptual quality of feature visualizations and robustness to targeted attributional attacks, we utilize different backbones (VGG-16, Resnet-18, InceptionV1) for the discriminative models. All the backbones are frozen during training the classifiers for the first 10 epochs, before fine-tuning the whole models for another 7 epochs.

Datasets. We validate our approach on two real-image datasets: (1) the Oxford-IIIT Pet Dataset (Parkhi et al., 2012), which compromises pet images of 37 categories, making it well-suited for fine-grained classification tasks. The images vary in pose and lighting, making it also apt for the task of examining how VAEs capture and represent the inherent properties of the underlying data manifold. (2) the Oxford 102 Flower Dataset (Nilsback & Zisserman, 2008), which is also used for fine-grained recognition tasks. It compromises 102 different flower categories, with a large variation in scale, pose, color, and background, reflecting the intricacies in real-world settings. This motivates our use of this dataset, as we believe that VAEs can effectively capture a data manifold with such complexity and intra-class variability.

To ensure a fair comparison of our method with IG, we need to use the standard black and white baseline images. Our goal is to generate geodesic paths of attribution on the manifold, linking these baseline images to the target images. However, due to the variational approximation in VAEs and their inherent noise, reconstructions of purely black or white images is not feasible. To incentivise the VAEs to produce cleaner black and white baseline reconstructions, we augment all datasets with black and white images during the training phase of the VAEs. This ensures VAEs can handle such plain images.

Baseline Methods. We compare our method against the following alternatives:

- **Saliency** (Simonyan et al., 2014). The simplest method for capturing sensitivity of the model to changes in the input.
- **Input** $\times$ **Gradients** (Shrikumar et al., 2017). A method proposed to add sharpness to the sensitivity maps.
- **Guided Backpropagation** (Springenberg et al., 2015). It uses a modification to gradient calculations in ReLU-based classifiers by only backpropagating non-negative gradients.
- **IG** (Sundararajan et al., 2017). The original path-based method that uses the linear path in the image space.
- **Smooth IG** (Smilkov et al., 2017): It reduces the noise in IG by averaging feature attributions from noisy images.
- **EIG** (Jha et al., 2020). In this approach, the manifold is conceptualized as a flat, Euclidean space, and it generates feature maps along straight linear paths in the latent space.

4.1. Geodesic Paths of Attribution

In curved manifolds, geodesics are the shortest paths that represent the most seamless transitions possible between two points. When applying this to latent representation of images, it translates to a smooth interpolation between two images, staying faithful to the structure of the data manifold. This means that the transformed geodesic path yields a series of realistic images with minimal perceptible differences between successive interpolations. In contrast, a linear interpolation path, based on Euclidean geometry, tends to deviate outside the data manifold, violating the curved nature of the underlying space.

The linear path in IG, $\gamma = x' + t(x - x')$, transforms all pixels with the same scale, which implies a complete independence amongst pixels, violating real-world settings. The progression of pixels along a smooth path on the data manifold, however, is a complex and interdependent process. This intricacy stems from the interconnected nature of pix-

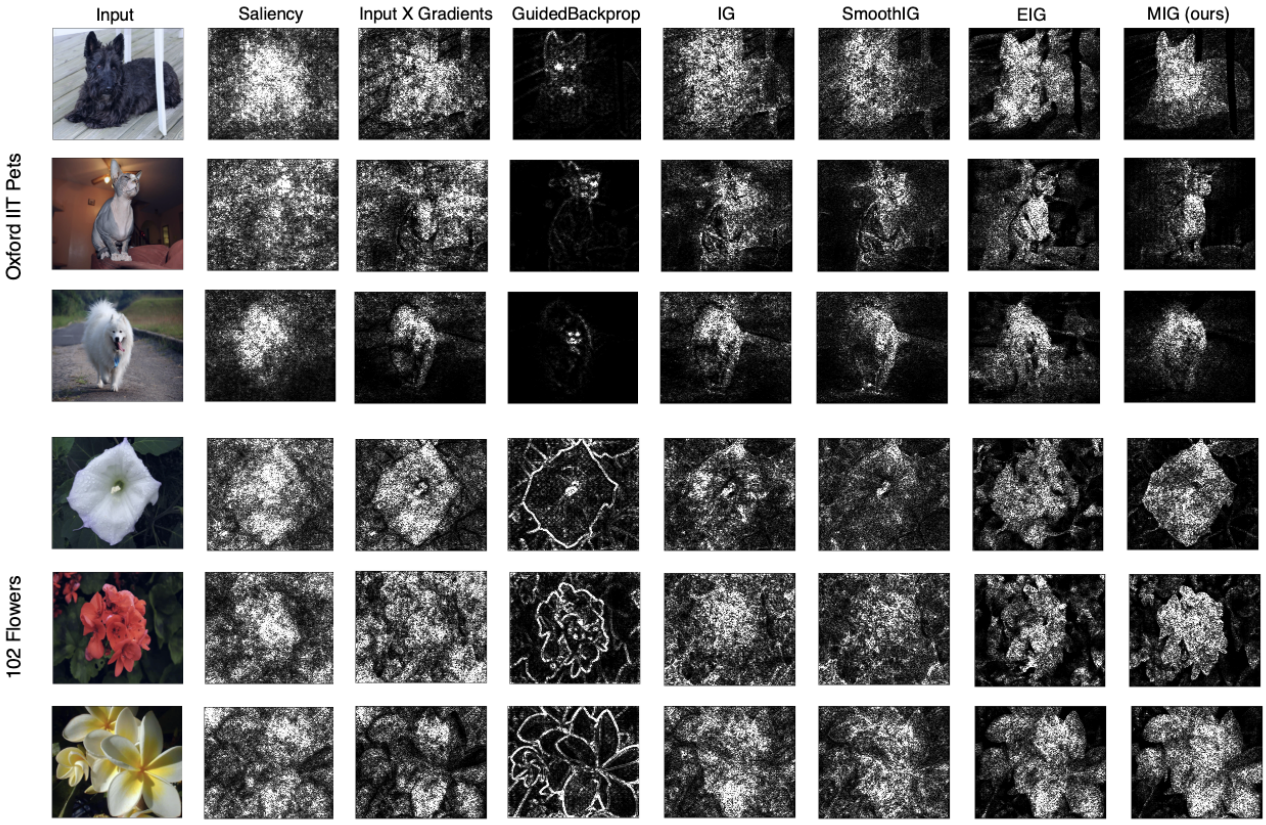

Figure 4. Comparison of Feature Attribution Methods. Presented are feature maps from various methods—Saliency, Gradients $\times$ Input, GuidedBackprop, IG, Smooth IG, EIG, as well as our proposed MIG. As shown, MIG addresses the IG's noise limitation and surpasses other methods, producing distinctly clearer and perceptually more aligned visualizations. In the last row, the similarity between EIG and MIG indicates that the path of attribution in MIG passes through a nearly flat region on the data manifold, and hence a linear interpolation path employed by EIG can closely approximate the mapped geodesic path in MIG for this particular image.

els, where each pixel does not exist in isolation. The essence lies in the idea that "*the whole is greater than the sum of its parts*," as the transitions are shaped by the influence of neighboring pixels. In Figure 3(a), we can see that superpixels on the smooth path surrounding dog's nose start to arise first, and pixels affect relevant pixels as they progress, until we reach the full target image. It is also worth highlighting that specific feature structure (the nose in this case) can emerge and evolve differently from other features, which means that progression is curved and context-dependent. This curvature in the data manifold is influenced by the distribution of intensities, colors, gradients within the image, and structural properties in the image, making the mapped geodesic path a complex function of all those influential elements and variations. In this sense, MIG along smooth paths on the manifold can capture feature interactions, which is a major limitation of the linear path in IG.

From the perspective of the model output, shown in Figure 3(b), mapped geodesics avoid the wide saturation region, where the model's response becomes less sensitive to changes in input features. As gradients saturate in these regions, the resulting attributions are predominantly influenced by noise. The linear path in IG is a significant contributing factor to the issue of saturation-induced noise (Kapishnikov et al., 2021; Miglani et al., 2020). This arises because the linear path inherently crosses through a wide saturation region. The essence of the problem lies in the nature of linear interpolation, which introduces discriminative image features early in the path. As a result, the model's response escalates rapidly at the beginning of the linear path and then quickly plateaus. The corresponding feature maps to the two paths of attribution are shown in Figure 3(c).

4.2. Perceptual Attribution Maps along the Geodesics

The underlying data manifold provides the natural structure for the model to learn. When model gradients are aligned

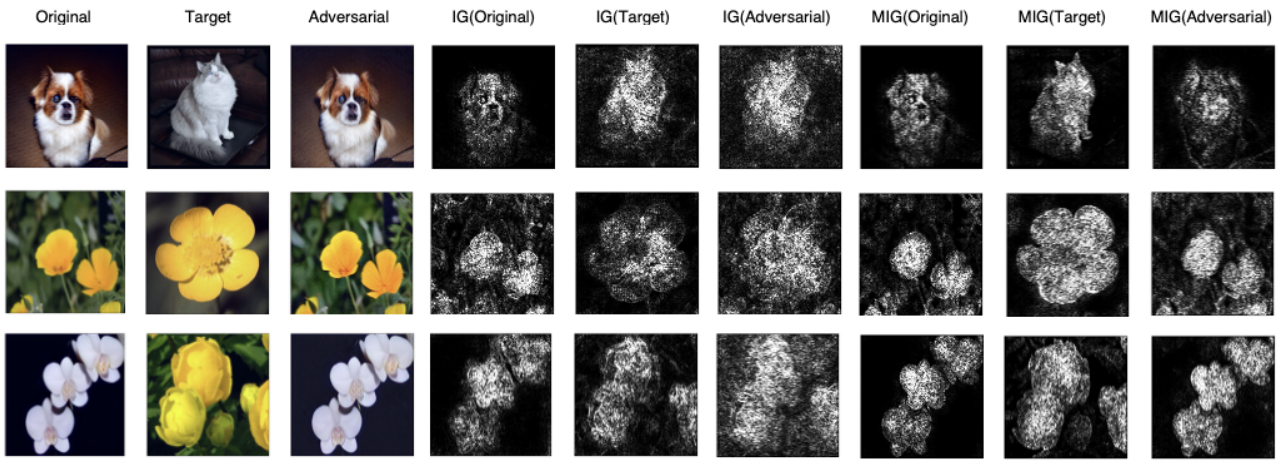

Figure 5. MIG vs. IG under targeted attributional attacks. The figure displays examples of an original input image alongside a target image and an adversarial attributional attack designed to exploit the IG's linear path of attribution. IG's vulnerability is evident as it generates adversarial feature maps that erroneously mimic the target maps. MIG maintains perceptually consistent and noise-resistant feature visualizations for the adversarial examples, closely resembling those of the original input. Each row was generated based on a different classifier's backbone, VGG-16, ResNet18, InceptionV1, respectively.

with this image manifold, it ensures that gradients are perceptually relevant. Geodesics come into play to enforce integrated gradients to be aligned with the manifold, emphasising features that are consistent with human perception. MIG, utilising transformed geodesic paths that conform to the curved natural structure of the data manifold, is essentially an accumulation of PAGs. In Figure 4, we show feature attribution maps generated by MIG (last column) as compared to other methods. In all examples, MIG clearly outperforms others as it produces perceptually consistent attributions with minimal noise. In the last row in Figure 4, one can see a certain degree of similarity between EIG and MIG feature maps. This is because, for this specific image, the linear path employed by EIG is an approximation to the geodesic due to the low curvature of the latent manifold along the geodesic from the baseline to the target.

4.3. Robustness to Targeted Attributional Attacks

Adversarial attributional attacks exploit model's sensitivity to imperceptible perturbations, which are usually off-manifold. Aligning the model's gradients with the data manifold, provides inherent robustness to such perturbations. The linear path of attribution in IG adds up to this vulnerability as it results in interpolants on the path being positioned outside the manifold. This shift of interpolants away from the manifold can potentially move them into regions of adversarial examples, leading to the accumulation of irregular gradients. VAEs possess denoising capabilities, allowing them to denoise input, including noisy or adversarial examples. This process can be seen as projecting noisy or adversarial inputs closer to the manifold of normal data. MIG, by adhering to the data manifold, increases robustness. Figure 5 shows how MIG feature maps are more robust to targeted attributional attacks as compared to IG. In this case, an input image is manipulated to compose an adversarial example. This generates attributions similar to an arbitrary target image, while maintaining the same output class as the original input. Unlike employing adversarial training to build robust classifiers that exhibit PAGs – a process that is computationally intensive and often detrimental to the model's performance (Tsipras et al., 2018) – our approach focuses on learning the data manifold and aligning model gradients with the intrinsic geometry. This approach effectively generates robust and perceptually aligned feature visualizations, achieving this without sacrificing the accuracy of the classifier, though it introduces some additional complexity as highlighted in Appendix C.

5. Quantitative Analysis

To assess our method (MIG), we use robust metrics from Yeh et al. (2019) and Wang et al. (2004) to evaluate the generated explanations. We compare MIG against IG, BlurIG, and SmoothIG (SIG), examining fidelity, sensitivity, and robustness through metrics such as infidelity, maximum sensitivity, and structural similarity.

5.1. Metrics

Explanation Infidelity (INFD) (Yeh et al., 2019). This measure is a robust variant of the *completeness* axiom. It

quantifies the degree to which an explanation misrepresents the model's sensitivity to input perturbations, assessing inaccuracies reflected in feature attributions. This contrasts with the completeness axiom that requires feature attributions to sum up to the total change in model output, without directly evaluating perturbation sensitivity. For a black-box function f, an explanation functional Φ, and a random vector ξ that characterizes significant perturbations of interest around input x, the explanation infidelity of Φ is defined as

$$\text{INFD}(f, x, \Phi) = \mathbb{E}_\xi \left[\langle \xi, \Phi(f, x) \rangle - (f(x) - f(x - \xi)) \right]^2,$$

where $\langle ., . \rangle$ denotes the Euclidean inner-product between vectors. One plausible way to determine ξ is by calculating the difference from a noisy baseline: $\xi = x - z_0$, where $z_0 = x_0 + \epsilon$. In this context, ϵ denotes a zero-mean random vector, such as $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

Maximum Sensitivity (SENS_{max}) (Yeh et al., 2019). The essence of max-sensitivity is to evaluate how sensitive an explanation is to perturbations in the input, with a particular focus on its maximum deviation. Given an input neighborhood radius r, maximum sensitivity is defined as

$$\text{SENS}_{\text{max}} = \max_{\|\delta\| \leq r} \|\Phi(f(x + \delta)) - \Phi(f(x))\|,$$

where $\|.\|$ is usual Euclidean norm of the vectorized input. It is worth noting that while reducing sensitivity might seem desirable to enhance robustness, doing so without careful consideration can lead to sub-optimal feature attributions. For instance, the intrinsic sensitivity of natural explanations—either due to the model's inherent sensitivity or the explainability method—suggests that some degree of sensitivity is unavoidable and, in fact, necessary to maintain the fidelity of the explanation to the model's behavior. However, it is possible to reduce the sensitivity of an explanation in a way that also lowers its infidelity. This dual benefit is significant, as it suggests that reliable explanations can be less sensitive, and also more accurate in representing the model's predictive behavior.

Structural Similarity Index (SSI) (Wang et al., 2004). We use this metric to assess MIG's robustness to IG-targeted attributional attacks. SSI can measure the similarity between attribution maps for the input and its adversarial version as it captures intricacies be evaluating the match in contrast levels, brightness (attribution scores), and structures within the attribution maps.

5.2. Results

Our method satisfies the axioms of path-based feature attribution while also ensuring reliability and faithfulness with robust metrics. Table 1 indicates that MIG achieves the lowest maximum sensitivity to input perturbations and significantly enhances faithfulness compared to other methods. This is viewed through the data-manifold perspective, where perturbations typically displace the input from the manifold. Our geodesic path of attribution, aligning with the manifold's geometry, renders the feature attribution maps less sensitive to perturbations and more faithful to the model's output behavior.

Datasets	Oxford IIT Pets		102 Flowers	
Methods	SENS_{max}	INFD	SENS_{max}	INFD
IG	0.87	7.65	0.74	15.26
BlurIG	0.75	6.41	0.67	12.19
SIG	0.42	4.30	0.38	9.81
MIG(ours)	**0.17**	**1.86**	**0.21**	**3.46**

Table 1. Sensitivity and Infidelity of Feature Attributions. Here, lower values signify better quality. Our method, MIG, achieves the highest quality among the alternatives.

Datasets	Oxford IIT Pets			102 Flowers		
Backbones	SSI			SSI		
	IG	SIG	MIG	IG	SIG	MIG
VGG-16	0.43	0.64	**0.87**	0.35	0.57	**0.74**
ResNet18	0.48	0.69	**0.91**	0.41	0.63	**0.86**
InceptionV1	0.36	0.61	**0.86**	0.36	0.52	**0.71**

Table 2. Structural Similarity Under IG-targeted Attributional Attacks. Here, higher values signify more robustness. Our method, MIG, achieves the greatest robustness among the alternatives.

Table 2 shows that MIG scores the highest SSI, indicating it preserves the structure and relevancy of attribution maps under targeted attributional attacks. Unlike IG, which uses a noninformative linear path accumulating adversarial gradients, and SIG, which averages multiple linear paths potentially smoothing feature maps for added robustness, MIG avoids these linear paths. Instead, it uses a single geodesic path that exhibits a higher degree of robustness.

6. Conclusion

Our work introduces MIG as a solution to the reliability issues in IG for deep learning models. MIG leverages geodesics on the latent manifold to provide smoother interpolations between images, capturing the non-linear nature of image manifolds. In contrast to the linear path in IG, MIG captures pixel interactions more realistically, reducing noise in feature attributions. Additionally, MIG enhances model's robustness against targeted attributional attacks by aligning gradients with the data manifold. Our experiments validate the effectiveness of MIG, offering perceptually aligned explanations and promising "safer" applications in domains requiring enhanced model interpretability and reliability, e.g., critical sectors such as healthcare.

Acknowledgements

This research was partially supported by the Australian Research Council through an Industrial Transformation Training Centre for Information Resilience (IC200100022). Quan Nguyen is supported by a NHMRC Investigator Grant (GNT2008928).

Impact Statement

This paper presents MIG, an approach that can potentially be applied to improve interpretability in a broad range of high-risk neural network applications. It complements the existing discussions on Explainable AI from a data-manifold perspective, and bridges the gap between reliability and effectiveness of gradient-based explainability techniques. A direct positive impact of the proposed method is to enhance transparency and accelerate adoption of complex vision models for medical imaging. However, while we enhance robustness to targeted attributional attacks, potentially new adversarial attacks can target our method and exploit vulnerabilities that we did not address.

References

Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent Space Oddity: on the Curvature of Deep Generative Models. February 2018.

Arvanitidis, G., Hauberg, S., Hennig, P., and Schober, M. Fast and Robust Shortest Paths on Manifolds Learned from Data. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pp. 1506–1515. PMLR, April 2019. ISSN: 2640-3498.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, July 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130140. Publisher: Public Library of Science.

Baniecki, H., Kretowicz, W., and Biecek, P. Fooling Partial Dependence via Data Poisoning. In Amini, M.-R., Canu, S., Fischer, A., Guns, T., Kralj Novak, P., and Tsoumakas, G. (eds.), *Machine Learning and Knowledge Discovery in Databases*, volume 13715, pp. 121–136. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-26408-5 978-3-031-26409-2. doi: 10.1007/978-3-031-26409-2_8. Series Title: Lecture Notes in Computer Science.

Bengio, Y., Courville, A., and Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50.

Bordt, S., Upadhyay, U., Akata, Z., and Von Luxburg, U. The Manifold Hypothesis for Gradient-Based Explanations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3697–3702, Vancouver, BC, Canada, June 2023. IEEE. ISBN 9798350302493. doi: 10.1109/CVPRW59228.2023.00378.

Brahma, P. P., Wu, D., and She, Y. Why Deep Learning Works: A Manifold Disentanglement Perspective. *IEEE Transactions on Neural Networks and Learning Systems*, 27(10):1997–2008, October 2016. ISSN 2162-237X, 2162-2388. doi: 10.1109/TNNLS.2015.2496947.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-VAE, April 2018. arXiv:1804.03599 [cs, stat].

Chalasani, P., Chen, J., Chowdhury, A. R., Wu, X., and Jha, S. Concise Explanations of Neural Networks using Adversarial Training. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1383–1391. PMLR, November 2020. ISSN: 2640-3498.

Chen, J., Wu, X., Rastogi, V., Liang, Y., and Jha, S. Robust Attribution Regularization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Cho, S., Lee, J., and Kim, D. Hyperbolic VAE via Latent Gaussian Distributions, October 2023. arXiv:2209.15217 [cs].

Cristovao, P., Nakada, H., Tanimura, Y., and Asoh, H. Generating In-Between Images Through Learned Latent Space Representation Using Variational Autoencoders. *IEEE Access*, 8:149456–149467, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.3016313.

Davidson, T. R., Falorsi, L., Cao, N. D., Kipf, T., and Tomczak, J. M. Hyperspherical Variational Auto-Encoders. 2018.

Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., and Kessel, P. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S.-I. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3(7):620–631, July 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00343-w. Number: 7 Publisher: Nature Publishing Group.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, February 2016. ISSN 0894-0347, 1088-6834. doi: 10.1090/jams/852.

Fong, R. C. and Vedaldi, A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, Venice, October 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.371.

Ganz, R., Kawar, B., and Elad, M. Do Perceptually Aligned Gradients Imply Adversarial Robustness?, August 2023. arXiv:2207.11378 [cs].

Ghorbani, A., Abid, A., and Zou, J. Interpretation of Neural Networks Is Fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33013681. Number: 01.

Heo, J., Joo, S., and Moon, T. Fooling Neural Network Interpretations via Adversarial Model Manipulation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. November 2016.

Hinton, G. E. and Zemel, R. Autoencoders, Minimum Description Length and Helmholtz Free Energy. In *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1993.

Hou, X., Shen, L., Sun, K., and Qiu, G. Deep Feature Consistent Variational Autoencoder, October 2016. arXiv:1610.00291 [cs].

Ivankay, A., Girardi, I., Marchiori, C., and Frossard, P. FAR: A General Framework for Attributional Robustness, March 2022. arXiv:2010.07393 [cs].

Jha, A., K. Aicher, J., R. Gazzara, M., Singh, D., and Barash, Y. Enhanced Integrated Gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome Biology*, 21(1):149, June 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02055-7.

Jost, J. *Riemannian geometry and geometric analysis*. Springer, 7 edition, 2017.

Kapishnikov, A., Venugopalan, S., Avci, B., Wedin, B., Terry, M., and Bolukbasi, T. Guided Integrated Gradients: an Adaptive Path Method for Removing Noise. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5048–5056, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.00501.

Kaur, S., Cohen, J., and Lipton, Z. C. Are Perceptually-Aligned Gradients a General Property of Robust Classifiers?, October 2019. arXiv:1910.08640 [cs, stat].

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes, December 2022. arXiv:1312.6114 [cs, stat].

Laberge, G., Aïvodji, U., Hara, S., Marchand, M., and Khomh, F. Fooling SHAP with Stealthily Biased Sampling. September 2022.

Lee, J. M. *Introduction to Smooth Manifolds*. Springer, 2 edition, 2014.

Lee, J. M. *Introduction to Riemannian manifolds*. Springer, 2 edition, 2018.

Lundberg, S. M. and Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Lundstrom, D., Huang, T., and Razaviyayn, M. A rigorous study of integrated gradients method and extensions to internal neuron attributions, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. February 2018.

Mathieu, E., Le Lan, C., Maddison, C. J., Tomioka, R., and Teh, Y. W. Continuous Hierarchical Representations with Poincaré Variational Auto-Encoders. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Miglani, V., Kokhlikyan, N., Alsallakh, B., Martin, M., and Reblitz-Richardson, O. Investigating Saturation Effects in Integrated Gradients, October 2020. arXiv:2010.12697 [cs].

Miolane, N. and Holmes, S. Learning Weighted Submanifolds With Variational Autoencoders and Riemannian Variational Autoencoders. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14491–14499, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.01451.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes, 2008.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, 1999.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. The oxford-iiit pet dataset, 2012.

Ran, X., Xu, M., Mei, L., Xu, Q., and Liu, Q. Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation. *Neural Networks*, 145:199–208, January 2022. ISSN 0893-6080. doi: 10.1016/j.neunet.2021.10.020.

Ribeiro, M. T., Singh, S., and 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*, pp. 1135–1144, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

Shah, H., Jain, P., and Netrapalli, P. Do Input Gradients Highlight Discriminative Features?, October 2021. arXiv:2102.12781 [cs, stat].

Shao, H., Kumar, A., and Fletcher, P. T. The Riemannian Geometry of Deep Generative Models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 428–4288, Salt Lake City, UT, USA, June 2018. IEEE. ISBN 978-1-5386-6100-0. doi: 10.1109/CVPRW.2018.00071.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3145–3153. PMLR, July 2017. ISSN: 2640-3498.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014.

Skopek, O., Ganea, O.-E., and Bécigneul, G. Mixed-curvature Variational Autoencoders. September 2019.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 180–186, New York, NY, USA, February 2020. Association for Computing Machinery. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375830.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. SmoothGrad: removing noise by adding noise, June 2017. arXiv:1706.03825 [cs, stat].

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for Simplicity: The All Convolutional Net, April 2015. arXiv:1412.6806 [cs].

Srinivas, S., Bordt, S., and Lakkaraju, H. Which Models have Perceptually-Aligned Gradients? An Explanation via Off-Manifold Robustness. November 2023.

Sturmfels, P., Lundberg, S., and Lee, S.-I. Visualizing the impact of feature attribution baselines. *Distill*, 2020. doi: 10.23915/distill.00022. https://distill.pub/2020/attribution-baselines.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Tifrea, A., Bécigneul, G., and Ganea, O.-E. Poincar\'e GloVe: Hyperbolic Word Embeddings, November 2018. arXiv:1810.06546 [cs].

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness May Be at Odds with Accuracy. September 2018.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pp. 1096–1103, Helsinki, Finland, 2008. ACM Press. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390294.

Wang, F. and Kong, A. W.-K. Exploiting the Relationship Between Kendall's Rank Correlation and Cosine Similarity for Attribution Protection, September 2022. arXiv:2205.07279 [cs].

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13 (4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Xu, S., Venugopalan, S., and Sundararajan, M. Attribution in Scale and Space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9677–9686, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00970.

Yang, R., Wang, B., and Bilgic, M. IDGI: A Framework to Eliminate Explanation Noise from Integrated Gradients. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23725–23734, Vancouver, BC, Canada, June 2023. IEEE. ISBN 9798350301298. doi: 10.1109/CVPR52729.2023.02272.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., and Ravikumar, P. On the (In)fidelity and Sensitivity for

Explanations, November 2019. URL http://arxiv.org/abs/1901.09392. arXiv:1901.09392 [cs, stat].

Zhu, J.-Y., Krähenbühl, P., Shechtman, E., and Efros, A. A. Generative Visual Manipulation on the Natural Image Manifold. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pp. 597–613, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46454-1. doi: 10.1007/978-3-319-46454-1_36.

A. Further Details on Computing Geodesics

For the sake of self-containment, we briefly review the approach taken by Shao et al. (2018, Algorithm 1) to compute the geodesics on the manifold. Subsequently, we will outline our modifications in this work.

For T time steps, $0 = t_0 < t_1 < \ldots < t_{T-1} < t_T = 1$, and discrete time interval $\delta t = (t_{i+1} - t_i) = 1/T$, we consider an approximate discretization of the curve $\gamma(t)$ on the latent manifold, $\mathcal{M}$, as $z_0, z_1, \ldots, z_T$, i.e., $z_i = \gamma(t_i)$. The smooth map g now gives a discrete path on the data manifold as $g(z_0), g(z_1), \ldots, g(z_T)$. Using forward finite differences, we get an approximation to the velocity of the curve $g(\gamma(t))$ at t_i as

$$\frac{dg(\gamma(t))}{dt}\,|_{t=t_i} \approx \frac{g(\gamma(t_{i+1})) - g(\gamma(t_i))}{t_{i+1} - t_i} = \frac{g(z_{i+1}) - g(z_i)}{\delta t}.$$

Now, the discrete analog to the energy functional (1) is given by

$$E(\mathbf{z}) = \frac{1}{2}\sum_{t=0}^{T-1} \frac{1}{\delta t} \left\| g(z_{i+1}) - g(z_i) \right\|^2, \quad \text{where} \quad \mathbf{z} = \begin{bmatrix} z_0 & z_1 & \ldots & z_T \end{bmatrix}.$$

Fixing z_0 and z_T as the specific start and end points for the geodesic path, we aim to minimize this discrete geodesic energy through gradient descent applied to the intermediate points on the curve, $z_1, ..., z_{T-1}$. The gradient with respect to z_i is thus

$$\frac{\partial}{\partial z_i} E(\mathbf{z}) = \frac{1}{\delta t} J_g^T(z_i)(g(z_{i+1}) - 2g(z_i) + g(z_{i-1})).$$

Instead of utilizing J_g^T for calculating gradients, Shao et al. (2018) opted for the faster-to-compute Jacobian of the encoder, J_e, as they considered the former to be computationally expensive. The resulting modified gradient can then be written as

$$\eta_i = \frac{1}{\delta t} J_e(z_i)(g(z_{i+1}) - 2g(z_i) + g(z_{i-1})). \tag{5}$$

However, to produce more faithful geodesics, we find that using the exact decoder is crucial. To achieve this, instead of forming the Jacobian explicitly, we access it only through Jacobian-vector products. For this, we reorganize the original gradient as

$$\frac{\partial}{\partial z_i} E(\mathbf{z}) = \frac{1}{\delta t} \left(\left(g(z_{i+1}) - 2g(z_i) + g(z_{i-1})\right)^T J_g(z_i) \right)^T \tag{6}$$

This reorganization enables the computation using vector-Jacobian product, which is more computationally efficient than explicitly computing the Jacobian matrix J_g. We use (6) in place of the modified gradient (5) in Algorithm 1. By doing this, we do not compromise the use of the generator function, which is the core surface model in our work. The resulting modified geodesic path algorithm is given in Algorithm 1.

Algorithm 1 Geodesic Path

input Two points, $z_0, z_T \in \mathcal{Z}$

$\mathbf{z}^{(0)} = \{z_i^{(0)}\}_{i=0}^T$ as linear interpolation between z_0 and z_T

for $k = 0, 1, \ldots$ **do**

  for $i \in \{1, \ldots, T-1\}$ **do**

    Compute the gradient $\frac{\partial}{\partial z_i} E(\mathbf{z}^{(k)})$ using (6)

    $z_i^{(k+1)} = z_i^{(k)} - \alpha^{(k)} \frac{\partial}{\partial z_i} E(\mathbf{z}^{(k)})$

  end for

end for

output Discrete geodesic path $z_0, z_1, \ldots, z_T \in \mathcal{Z}$

For our experiment in Section 4, we employ backtracking line search with Armijo-Goldstein condition (Nocedal & Wright, 1999), to determine the appropriate step size, $\alpha^{(k)}$, for each gradient descent step. The outer loop is terminated after a maximum of 300 iterations or if $\left|\Delta E^{(k)} - \Delta E^{(k-1)}\right| \leq 0.001 \Delta E^{(k)}$ where $\Delta E^{(k)} = \sum_i \|\partial E(\mathbf{z}^{(k)})/\partial z_i\|^2$.

B. Details on the VAE Architecture

The VAE used in our experiments, as specified in the code repository, relies intensively on residual blocks and certain activations that allows the decoder to impose a valid Riemannian metric. Table 3 shows details of the architectural choices in the VAE model.

Module	Layer	Output Shape	Details
Encoder			
Input	-	$3 \times 192 \times 192$	-
Conv In	Conv2d	$64 \times 192 \times 192$	Kernel: 3, Stride: 1, Padding: 1, Activation: SiLU
ResDown1	Conv2d x2, BN x2	$128 \times 96 \times 96$	Conv1: Kernel: 3, Stride: 2, Padding: 1 Conv2: Kernel: 3, Stride: 1, Padding: 1 Activation: SiLU
ResDown2	Conv2d x2, BN x2	$256 \times 48 \times 48$	Same as above
ResDown3	Conv2d x2, BN x2	$512 \times 24 \times 24$	Same as above
ResDown4	Conv2d x2, BN x2	$512 \times 12 \times 12$	Same as above
ResBlock	Conv2d x2, BN x2	$512 \times 12 \times 12$	Kernel: 3, Stride: 1, Padding: 1, Activation: SiLU
Mu	Conv2d	$64 \times 12 \times 12$	Kernel: 1, Stride: 1, Padding: 0
Log Var	Conv2d	$64 \times 12 \times 12$	Kernel: 1, Stride: 1, Padding: 0
Decoder			
Conv In	Conv2d	$512 \times 12 \times 12$	Kernel: 1, Stride: 1, Padding: 0, Activation: ELU
ResUp1	Upsample, Conv2d x2, BN x2	$512 \times 24 \times 24$	Upsample: Scale: 2 Conv1: Kernel: 3, Stride: 1, Padding: 1 Conv2: Kernel: 3, Stride: 1, Padding: 1 Activation: ELU
ResUp2	Upsample, Conv2d x2, BN x2	$256 \times 48 \times 48$	Same as above
ResUp3	Upsample, Conv2d x2, BN x2	$128 \times 96 \times 96$	Same as above
ResUp4	Upsample, Conv2d x2, BN x2	$64 \times 192 \times 192$	Same as above
Conv Out	Conv2d	$3 \times 192 \times 192$	Kernel: 3, Stride: 1, Padding: 1, Activation: Tanh

Table 3. Detailed architecture of the VAE, showcasing the configurations of layers within the Encoder and Decoder modules.

C. Limitations of our Approach

While our approach bridges the gap between the robustness of attribution maps under adversarial conditions and the perceptual alignment and intuitiveness of explanations, it requires the use of VAEs to capture the underlying data manifold. Two sources of complexity arise from this setting: the training of VAEs and, subsequently, the geodesic computations necessary to generate the path of attribution in MIG. Future directions could aim to capture the Riemannian structure or specify the Riemannian metric needed to compute geodesics in ways that are less costly than using the VAE setup.