Counterfactual Explanations on Robust Perceptual Geodesics

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

[1] ARC Training Centre for Information Resilience (CIRES), [2] The University of Queensland, [3] QIMR Berghofer Medical Research Institute

Introduction

Motivation

Counterfactual explanations are framed as finding the smallest change that flips a model's prediction. In high-dimensional vision settings, however, "smallest" is fundamentally ambiguous: it is defined by the chosen distance metric, which in turn induces the geometry of the optimization problem.

Fig 1. Interpolations paths under different latent space geometries

Contribution

We introduce **Perceptual Counterfactual Geodesics (PCG)**, a geometry-aware approach that defines counterfactual paths as geodesics under a robust perceptual Riemannian metric in latent space. This formulation brings together manifold structure, perceptual alignment, and robustness, producing counterfactual trajectories that are smoother, semantically more faithful, and less prone to metric exploitation than prior approaches.

Perceptual Counterfactual Geodesics

Metric Construction: G_R aggregates pullbacks from robust model layers (h_k) to align with robust human perception

$$G_R(x) = \sum_{k=1}^{K} w_k \, J_{h_k}(x)^\top J_{h_k}(x), \quad w_k = \frac{1}{N_k}$$

Latent Geometry: The induced metric G_Z favours smooth, on-manifold, and semantically robust trajectories $\gamma(t)$

$$E\big(g(\gamma)\big) = \frac{1}{2}\int_0^1 \gamma'(t)^\top G_Z\big(\gamma(t)\big)\gamma'(t)dt$$

Proposed Method

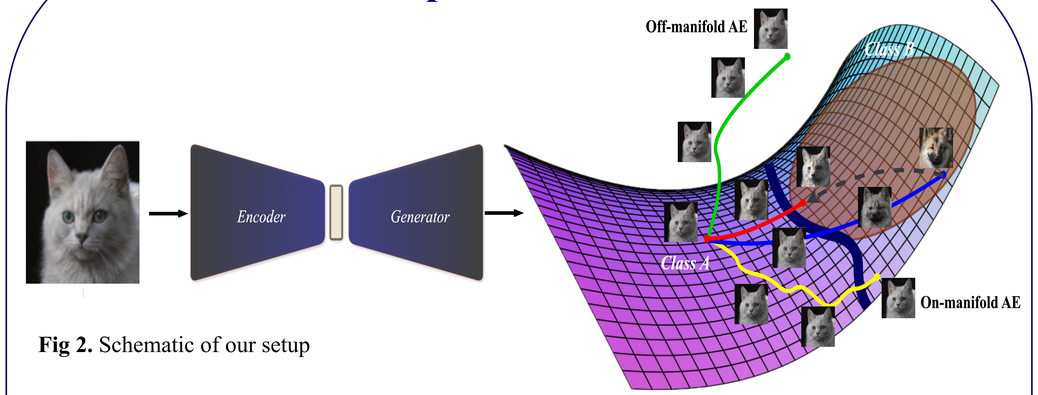

Fig 2. Schematic of our setup

Path Optimization: We minimize discretized energy across T waypoints to find a robust geodesic

$$E_{robust}(z) = \frac{1}{2}\sum_{i=0}^{T-1}\sum_{k=1}^{K}\frac{w_k}{\delta t}\,\|h_k\big(g(z_{i+1})\big) - h_k\big(g(z_i)\big)\|_2^2$$

Two-stage Optimization: The final objective balances the geometric energy with a classification loss ℓ to maintain the target class y' while staying on the robust manifold

$$\mathcal{L}(z) = E_{robust}(z) + \lambda \cdot \ell\big(f\big(g(z_T)\big), y'\big)$$

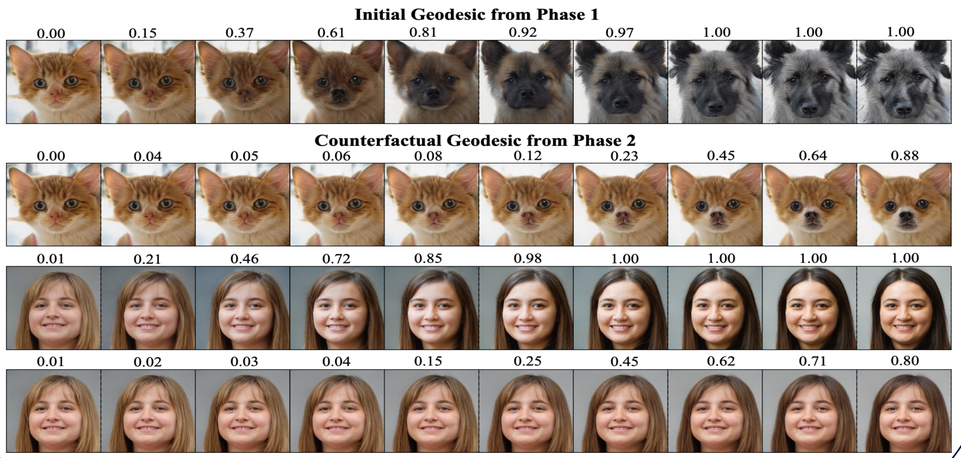

Fig 3. PCG stages**:** Initial Geodesics (Rows 1, 3) vs. Phase 2 Refined Counterfactuals (Rows 2, 4)

Experiments

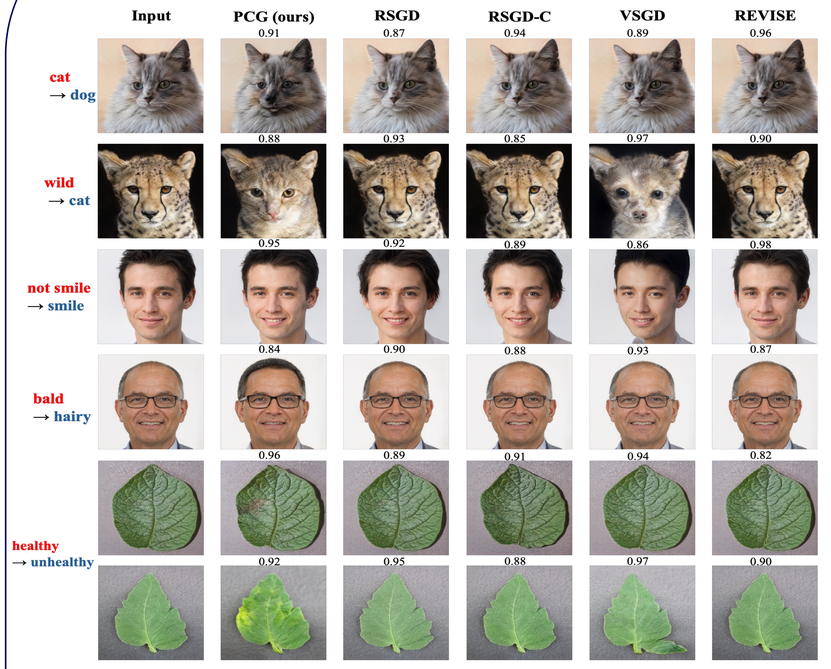

Fig 4. Qualitative comparison of counterfactuals across methods

	Realism ↓		Closeness ↓		Faithfulness ↑		Flip ↑
Method	**FID**	**R-FID**	**LPIPS**	**R-LPIPS**	**COUT**	**Mean SM**	**Rate**
REVISE	18.5	50.1	0.85	0.67	0.09	−0.48	**98%**
VSGD	23.5	46.7	0.93	0.79	0.10	−0.14	92%
RSGD	12.9	37.8	0.61	0.68	0.13	0.03	96%
RSGD-C	12.7	28.3	0.59	0.53	0.25	0.05	94%
PCG (ours)	**8.3**	**9.1**	**0.24**	**0.17**	**0.43**	**0.74**	95%

Table 1. Evaluation results across realism, closeness, faithfulness, and flip rate

Paper

Code