# Plausible Visual Counterfactual Explanations in Latent Space with Normalizing Flows

Łukasz Lenkiewicz[1 [0009−0006−6282−8649]⋆], Marcel Musiałek[1 [0009−0009−4964−0547]], Oleksii Furman[1 [0009−0001−2184−3096]], and Maciej Zięba[1 [0000−0003−4217−7712]]

Wrocław University of Science and Technology,
Wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław, Poland
lukasz.lenkiewicz@pwr.edu.pl

**Abstract.** Counterfactual explanations provide interpretable insights into classifier decisions by identifying minimal input modifications that alter predictions. While extensively studied for tabular data, visual counterfactuals present unique challenges requiring semantically meaningful changes rather than imperceptible perturbations. Current approaches predominantly employ diffusion models, GANs, and VAEs, while normalizing flows remain underexplored despite offering tractable likelihood computation. We introduce PLACE (Plausible LAtent Counterfactual Explanations), a method that leverages conditional normalizing flows for explicit density estimation in counterfactual generation. Operating in the latent space of a pre-trained autoencoder, PLACE optimizes a novel composite loss function balancing validity, proximity, and plausibility. The plausibility term directly maximizes log-likelihood under the target class distribution, enabled by the flow's tractable density computation. Experiments on CelebA and MNIST demonstrate that PLACE achieves competitive performance across multiple metrics while uniquely satisfying explicit plausibility constraints through substantially improved log density scores. Our method balances computational efficiency with multi-objective optimization, validating normalizing flows as an effective approach for probabilistically constrained visual counterfactual explanations.

**Keywords:** Machine Learning · Counterfactual Explanations · Explainable Artificial Intelligence · Computer Vision.

## 1 Introduction

Counterfactual explanations (CEs) [23] identify minimal perturbations to input instances that change a classifier's prediction, providing insight into learned decision boundaries. Effective CEs satisfy three properties: *minimality* (small perturbations), *realism* (plausible under the data distribution), and *actionability* (feasible to implement). While extensively studied for tabular data [5, 16], generating

⋆ Corresponding author

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_25

visual counterfactuals poses significant challenges. Unlike adversarial examples that achieve misclassification through imperceptible pixel-level noise, counterfactual explanations require semantically meaningful changes (e.g., adding a smile to a face or changing an object's color) while preserving class-irrelevant attributes such as background, pose, and identity.

This semantic requirement has driven the adoption of generative models for visual counterfactual generation. Current approaches predominantly employ diffusion models [7, 3], GANs [1, 14], and VAEs [21, 8]. Despite their proven effectiveness for density estimation, normalizing flows [17] remain underexplored for this task [6]. This is surprising given that flows offer a unique advantage: direct likelihood computation, allowing us to quantify the plausibility of a counterfactual as a member of the target class. This property is unavailable in other generative model families. Such probabilistic guarantees are particularly relevant in computational science applications where model interpretability must be paired with rigorous uncertainty quantification, for instance when explaining classifiers deployed in medical imaging [2], materials science [28], or computational biology [19], where stakeholders need to trust that an explanation is not only valid but also distributionally plausible.

We address this gap by introducing PLACE (**P**lausible **LA**tent **C**ounterfactual **E**xplanations), one of the first methods to utilize conditional normalizing flows for density estimation of generated counterfactuals. PLACE optimizes a novel composite loss that jointly enforces decision flipping, plausibility under the target class distribution, and proximity via L2 and LPIPS perceptual components. We evaluate on CelebA and MNIST, demonstrating competitive performance across multiple metrics against established baselines.

## 2 Related Works

*Normalizing Flows.* Normalizing flows enable exact density estimation through invertible transformations with tractable Jacobian determinants [17, 15]. However, flows applied directly in pixel space may capture local correlations rather than semantic content, motivating their use in learned embedding spaces. In the counterfactual domain, PPCEF [24] optimizes explicit density functions for tabular data and CeFlow [4] uses invertible flows for mixed-type features. PluGeN [25] employs flows to disentangle attributes in pre-trained latent spaces, which we build upon for conditional density estimation.

*Visual Counterfactual Explanations.* Wachter et al. [23] formalized CEs as an optimization balancing classifier loss and input distance, though without plausibility constraints. REVISE [8] addresses this by optimizing in a VAE's latent space, constraining the search to the learned data manifold but suffering from a plateau effect near decision boundaries. CLARITY [21] mitigates this by training classifier ensembles directly in the latent space, producing smoother boundaries at the cost of retraining. Other notable approaches include DiVE [18] for diverse counterfactuals, prototype-guided methods [12], and GAN-based approaches [20,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_25

14]. Diffusion-based methods such as DiME [7] and DVCE [3] achieve strong results but at higher computational cost. Recent work increasingly treats plausibility as an explicit constraint [22, 13], motivating our use of normalizing flows for direct density-based optimization.

## 3 Method

We propose PLACE, a method for generating counterfactual explanations that operates in the latent space of a pre-trained autoencoder enhanced with normalizing flows. Our approach addresses three key desiderata for visual counterfactuals: (i) *validity*, i.e., flipping the classifier's decision, (ii) *proximity*, i.e., remaining close to the original input, and (iii) *plausibility*, i.e., staying within high-probability regions of the learned data distribution.

### 3.1 Counterfactual Optimization

We consider a discriminative differentiable model $p_d(y|x)$ (e.g., a CNN classifier). Given an input image $x$ with latent representation $z = E(x)$ obtained via the encoder $E$, current class $y$, and target label $\tilde{y} \neq y$, we optimize a counterfactual embedding $z' \in \mathbb{R}^d$ in the $d$-dimensional latent space. The counterfactual image $\tilde{x}$ is generated as $\tilde{x} = G(z')$, where $G$ denotes the decoder. We minimize the following composite loss function:

$$\mathcal{L}(z') = \lambda_{\text{dec}}\mathcal{L}_{\text{decision}} + \lambda_{\text{dist}}\mathcal{L}_{\text{distance}} + \lambda_{\text{plaus}}\mathcal{L}_{\text{plausibility}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perceptual}}, \quad (1)$$

where $\lambda_{\text{dec}}$, $\lambda_{\text{dist}}$, $\lambda_{\text{plaus}}$, and $\lambda_{\text{perc}}$ control the relative importance of each term.

*Decision Loss.* Following Wielopolski et al. [24], we use a margin-based loss to ensure prediction flip:

$$\mathcal{L}_{\text{decision}}(\tilde{x}, \tilde{y}) = \max(0.5 + \epsilon - p_d(\tilde{y}|\tilde{x}), 0), \quad (2)$$

where $\epsilon > 0$ provides a safety margin beyond the decision boundary, ensuring robust flips that are stable under small perturbations.

*Distance Loss.* The distance loss penalizes large deviations from the original latent representation:

$$\mathcal{L}_{\text{distance}}(z, z') = \|z - z'\|_2, \quad (3)$$

promoting minimal-magnitude changes that preserve the original image's semantic content.

*Plausibility Loss.* The plausibility loss leverages the normalizing flow's exact density computation:

$$\mathcal{L}_{\text{plausibility}}(z', \tilde{y}) = -\log p_F(z'|\tilde{y}), \quad (4)$$

where $p_F(\cdot|\tilde{y})$ is the conditional normalizing flow model estimating the density for the target class $\tilde{y}$. This ensures counterfactuals lie within high-density regions of the target class distribution, preventing unrealistic examples.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_25

*Perceptual Loss.* The perceptual loss preserves high-level visual features:

$$\mathcal{L}_{\text{perceptual}}(x, \tilde{x}) = \text{LPIPS}(x, \tilde{x}), \tag{5}$$

where LPIPS [27] measures distance in deep feature space. This complements $\mathcal{L}_{\text{distance}}$ by enforcing proximity at a semantic level, ensuring that texture, structure, and identity remain consistent.

### 3.2 Architecture

Following PluGeN [25], we combine a deterministic convolutional autoencoder (encoder $E : \mathcal{X} \to \mathcal{Z}$, decoder $G : \mathcal{Z} \to \mathcal{X}$, $\mathcal{Z} \subset \mathbb{R}^d$) with an attribute-factorized normalizing flow $F : \mathcal{Z} \to \mathbb{R}^d$ that enables tractable conditional density computation in the latent space. A separately trained classifier $p_d(y|x)$ provides the decisions we aim to explain. All components remain frozen during counterfactual generation. Figure 1 illustrates the complete pipeline.

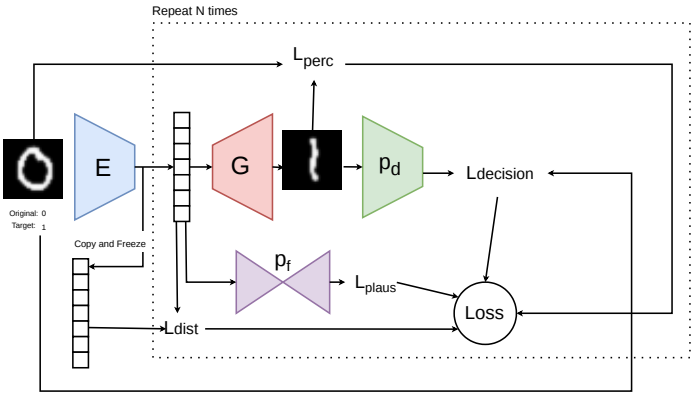


**Fig. 1.** Overview of the PLACE architecture. The input image is encoded into the latent space, and the embedding is iteratively optimized by minimizing a composite loss that combines decision, distance, plausibility, and perceptual terms. The decoder, classifier, and normalizing flow all remain frozen during optimization.

## 4 Experiments

We evaluate PLACE against established baselines on two benchmark datasets: CelebA [11] (smile classification, images resized to $256 \times 256$, standard train/test split) and MNIST [10] (digit-to-digit transformations such as 0→8 and 1→5,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_25

images resized to $32 \times 32$). We measure validity, LPIPS [27], FID [26], L2 distance, log density $\log p_F(z'|\tilde{y})$, and wall-clock time. We compare against CLARITY [21] and REVISE [8], two established latent-space counterfactual methods that share our autoencoder-based pipeline; a comparison with diffusion-based approaches [7, 3], which operate under fundamentally different generative paradigms, is left to future work.

### 4.1 Hyperparameters

All methods use Adam [9] with early stopping. For PLACE on CelebA: learning rate $5 \times 10^{-3}$, $\lambda_{\text{dec}} = 1000$, $\lambda_{\text{dist}} = 1.0$, $\lambda_{\text{plaus}} = 2.0$, $\epsilon = 0.05$. On MNIST: learning rate $10^{-2}$, $\lambda_{\text{dec}} = 10.0$, $\lambda_{\text{dist}} = 1.0$, $\lambda_{\text{plaus}} = 10.0$. REVISE and CLARITY baselines follow their original configurations. The pretrained classifier is ResNet-34 for CelebA (>95% accuracy) and a convolutional network for MNIST.

### 4.2 Results

Figure 2 illustrates the optimization process on CelebA, showing how PLACE progressively modifies discriminative features to achieve class change while preserving non-discriminative attributes.

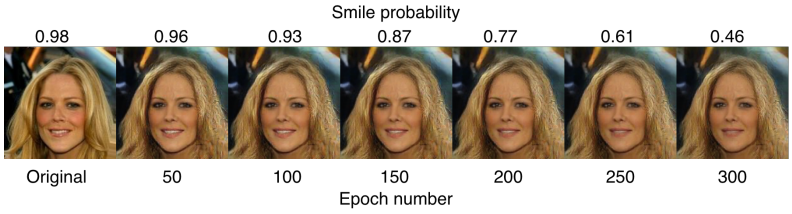


**Fig. 2.** Counterfactual generation on CelebA. PLACE progressively modifies smile-related features until the classifier's confidence for the original class drops below 50%, preserving identity and non-discriminative attributes.

*CelebA Results.* Table 1 presents results for smile classification counterfactuals. All methods achieve perfect validity. PLACE achieves the best LPIPS (0.21) and competitive log density (−1950.56), demonstrating effective balance between perceptual quality and plausibility. CLARITY achieves the best FID (105.01) and L2 (79.04) but requires the longest computation. REVISE is fastest but produces substantially worse perceptual quality (LPIPS 0.952).

*MNIST Results.* Table 2 shows results on MNIST. All methods achieve perfect validity. PLACE achieves substantially better log density (−120.66) compared to CLARITY (−134.11) and REVISE (−135.67), demonstrating the effectiveness

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_25

**Table 1.** Quantitative results on CelebA dataset. All methods achieve 100% validity.

| Method | Validity | LPIPS ↓ | FID ↓ | L2 ↓ | Log Density ↑ | Time (s) ↓ |
|---|---|---|---|---|---|---|
| CLARITY | 1.0 | 0.244 | **105.01** | **79.04** | -2210.19 | 10424.38 |
| REVISE | 1.0 | 0.952 | 442.67 | 265.37 | **-1909.05** | **4268.71** |
| Ours | 1.0 | **0.21** | 124.27 | 92.69 | -1950.56 | 7525.64 |

**Table 2.** Quantitative results on MNIST dataset. All methods achieve 100% validity.

| Method | Validity | L2 ↓ | Log Density ↑ | Time (s) ↓ |
|---|---|---|---|---|
| CLARITY | 1.0 | 13.87 | -134.11 | 41.13 |
| REVISE | 1.0 | **13.83** | -135.67 | **0.57** |
| Ours | 1.0 | 15.97 | **-120.66** | 14.39 |

of explicit density-based plausibility constraints, with slightly higher L2 distance (15.97 vs. ∼13.8) as a trade-off.

Overall, PLACE uniquely satisfies explicit plausibility constraints through normalizing flows while maintaining competitive proximity and perceptual quality. The substantial log density improvement validates the advantage of tractable likelihood computation for enforcing probabilistic plausibility, a property unavailable in baseline methods.

## 5 Conclusions and Future Work

We presented PLACE, a method for generating visual counterfactual explanations using conditional normalizing flows for explicit plausibility estimation in the latent space of a pre-trained autoencoder. Experiments on CelebA and MNIST demonstrate that PLACE achieves competitive performance across multiple metrics while uniquely satisfying explicit plausibility constraints, as evidenced by substantially improved log density scores. Our explicit density-based optimization provides direct control over plausibility, making PLACE particularly suitable for safety-critical applications requiring trustworthy explanations.

Several directions remain for future work. First, we plan to benchmark PLACE against diffusion-based counterfactual methods [7, 3] to provide a more comprehensive comparison across generative paradigms. Second, we intend to evaluate our framework on domain-specific datasets from computational science, such as medical imaging [2] and materials characterization [28], where plausibility-constrained explanations can directly support scientific discovery and clinical decision-making. Finally, scaling PLACE to higher-resolution images and exploring more expressive flow architectures may further improve the quality and applicability of the generated counterfactuals.

**Acknowledgments.** Oleksii Furman, Łukasz Lenkiewicz and Maciej Zięba's work was supported by the National Science Centre (Poland) Grant No. 2024/55/B/ST6/02100.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_25

## References


1. Atad, M., Dmytrenko, V., Li, Y., Zhang, X., Keicher, M., Kirschke, J., Wiestler, B., Khakzar, A., Navab, N.: Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan. arXiv preprint arXiv:2207.07553 (2022)
2. Atad, M., Schinz, D., Möller, H.K., Graf, R., Wiestler, B., Rueckert, D., Navab, N., Kirschke, J.S., Keicher, M.: Counterfactual explanations for medical image classification and regression using diffusion autoencoder. vol. abs/2408.01571 (2024). https://doi.org/10.48550/ARXIV.2408.01571
3. Augustin, M., Boreiko, V., Croce, F., Hein, M.: Diffusion visual counterfactual explanations. In: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022)
4. Duong, T.D., Li, Q., Xu, G.: Ceflow: A robust and efficient counterfactual explanation framework for tabular data using normalizing flows. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 133–144. Springer (2023)
5. Furman, O., Movsum-zada, U., Marszalek, P., Zięba, M., Śmieja, M.: Dicoflex: Model-agnostic diverse counterfactuals with flexible control. arXiv preprint arXiv:2505.23700 (2025)
6. Hvilshøj, F., Iosifidis, A., Assent, I.: Ecinn: efficient counterfactuals from invertible neural networks. arXiv preprint arXiv:2103.13701 (2021)
7. Jeanneret, G., Simon, L., Jurie, F.: Diffusion models for counterfactual explanations. In: Computer Vision - ACCV 2022 - 16th Asian Conference on Computer Vision, Macao, China, December 4-8, 2022, Proceedings, Part VII. Lecture Notes in Computer Science, vol. 13847, pp. 219–237. Springer (2022). https://doi.org/10.1007/978-3-031-26293-7_14
8. Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., Ghosh, J.: Towards realistic individual recourse and actionable explanations in black-box decision making systems. arXiv preprint arXiv:1907.09615 (2019)
9. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (2002)
11. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 3730–3738 (2015)
12. Looveren, A.V., Klaise, J.: Interpretable counterfactual explanations guided by prototypes. In: Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17, 2021, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12976, pp. 650–665. Springer (2021). https://doi.org/10.1007/978-3-030-86520-7_40
13. Melistas, T., Spyrou, N., Gkouti, N., Sanchez, P., Vlontzos, A., Panagakis, Y., Papanastasiou, G., Tsaftaris, S.A.: Benchmarking counterfactual image generation. In: Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024)
14. Mertes, S., Huber, T., Weitz, K., Heimerl, A., André, E.: Ganterfactual - counterfactual explanations for medical non-experts using generative adversarial learning. Frontiers Artif. Intell. **5**, 825565 (2022). https://doi.org/10.3389/FRAI.2022.825565


ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_25

15. Papamakarios, G., Murray, I., Pavlakou, T.: Masked autoregressive flow for density estimation. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 2338–2347 (2017)
16. Pawelczyk, M., Broelemann, K., Kasneci, G.: Learning model-agnostic counterfactual explanations for tabular data. In: Proceedings of the web conference 2020. pp. 3126–3132 (2020)
17. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)
18. Rodríguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I.H., Charlin, L., Vázquez, D.: Beyond trivial counterfactual explanations with diverse valuable explanations. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 1036–1045. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00109
19. Sapoval, N., Aghazadeh, A., Nute, M.G., Antunes, D.A., Balaji, A., Baraniuk, R., Barberan, C., Dannenfelser, R., Dun, C., Edrisi, M., et al.: Current progress and open challenges for applying deep learning across the biosciences. Nature Communications **13**(1), 1728 (2022)
20. Sauer, A., Geiger, A.: Counterfactual generative networks. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
21. Theobald, C., Pennerath, F., Conan-Guez, B., Couceiro, M., Napoli, A.: Clarity: an improved gradient method for producing quality visual counterfactual explanations. arXiv preprint arXiv:2211.15370 (2022)
22. Tsiourvas, A., Sun, W., Perakis, G.: Manifold-aligned counterfactual explanations for neural networks. In: International Conference on Artificial Intelligence and Statistics. pp. 3763–3771. PMLR (2024)
23. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
24. Wielopolski, P., Furman, O., Stefanowski, J., Zieba, M.: Probabilistically plausible counterfactual explanations with normalizing flows. In: ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024). Frontiers in Artificial Intelligence and Applications, vol. 392, pp. 954–961. IOS Press (2024)
25. Wolczyk, M., Proszewska, M., Maziarka, L., Zieba, M., Wielopolski, P., Kurczab, R., Smieja, M.: Plugen: Multi-label conditional generation from pre-trained models. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. pp. 8647–8656. AAAI Press (2022). https://doi.org/10.1609/AAAI.V36I8.20843
26. Yu, Y., Zhang, W., Deng, Y.: Frechet inception distance (fid) for evaluating gans. China University of Mining Technology Beijing Graduate School **3**(11) (2021)
27. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
28. Zhong, X., Gallagher, B., Liu, S., Kailkhura, B., Hiszpanski, A., Han, T.Y.J.: Explainable machine learning in materials science. npj computational materials **8**(1), 204 (2022)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_25