# Reliable Visual Counterfactual Explanation Using Conditional Flow Matching

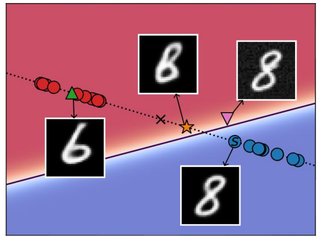
Zhuo Cao[1], Xuan Zhao[1*], Lena Krieger[1,2*], Hanno Scharr[1], Ira Assent[1,3]

[1]IAS-8, Forschungszentrum Jülich, Germany [2]Munich Center for Machine Learning (MCML), LMU Munich, Germany
[3]Aarhus University, Denmark

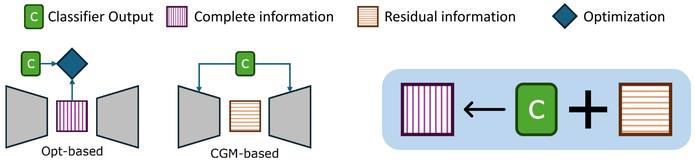
## TL; DR

We propose **LeapFactual**. It offers a flexible, **model-agnostic** mechanism for generating counterfactual explanations, capable of producing not only **high-quality** but also **reliable** results, even in the presence of **discrepancies between true and learned decision boundaries**.


## Counterfactual Explanations (CE)



Counterfactual Explanation (CE) seeks to make only **semantically meaningful modifications** to an input image in order to obtain a **similar image** with a **target label** prediction outcome.




### Optimization (Opt)-based Methods:

- ☹ Stop once the model's prediction matches the target label
- ☹ Require a differentiable classifier

### Conditional Generative Model (CGM)-based Methods:

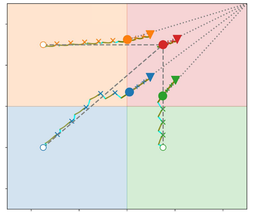
- ☹ Without control over boundary distance
- ☹ Need fine-tuning whenever the classifier changes

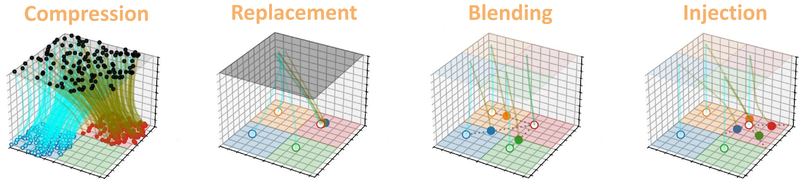
## LeapFactual

**Training Phase:** We demonstrate that the classifier's output is **disentangled** from the residual by training with our proposed Counterfactual Explanation–Conditional Flow Matching **(CE-CFM)** objective:

$$\mathcal{L}(\psi) := \mathbb{E}_{t,q(h),p_t(z|h)} \left\| v_\psi(t, z, c) - u_t(z|h) \right\|$$

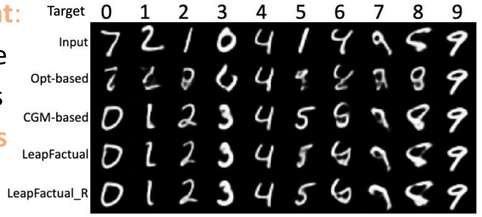
**Explaining Phase:** We propose two algorithms, LEAP and LeapFactual. Leveraging the flexibility of CFM, LEAP enables information replacement, **blending**, and **injection**. LeapFactual further integrates blending and injection leaps to generate **reliable** counterfactual explanations.



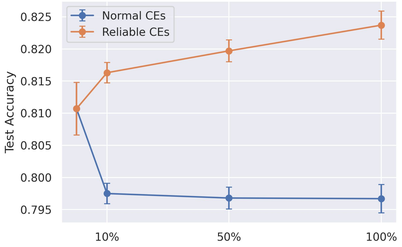



## Experiments

**Quantitative Assessment:** Our methods generate **realistic** counterfactuals by altering **class features** while **preserving style**.




**Model Improvement:** Model performance **drops** with **standard CE** blending, but both accuracy and AUC **improve** with **reliable CE** blending.




**Generalization:** We show that the proposed method is highly scalable to **high-res** images and **non-differentiable** classifiers.

# Reliable Visual Counterfactual Explanation Using Conditional Flow Matching

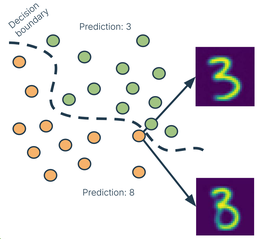
Zhuo Cao[1], Xuan Zhao[1*], Lena Krieger[1,2*], Hanno Scharr[1], Ira Assent[1,3]

[1]IAS-8, Forschungszentrum Jülich, Germany [2]Munich Center for Machine Learning (MCML), LMU Munich, Germany
[3]Aarhus University, Denmark

## TL; DR

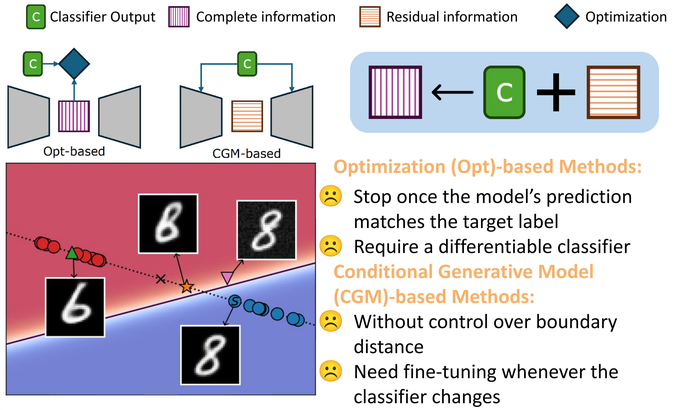
We propose **LeapFactual**. It offers a flexible, **model-agnostic** mechanism for generating counterfactual explanations, capable of producing not only **high-quality** but also **reliable** results, even in the presence of **discrepancies between true and learned decision boundaries**.

## What is Counterfactual Explanation (CE)




Counterfactual explanation seeks to make only **semantically meaningful modifications** to an input image in order to obtain a **similar image** with a **target label** prediction outcome. It exposes decision boundaries more clearly and can also be used to generate new, informative **training samples** that improve generalization and fairness.

## The Gap: Existing CE methods have limitations



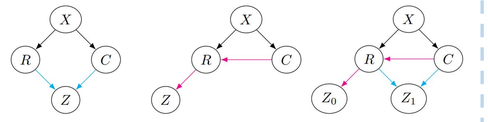

**Optimization (Opt)-based Methods:**

- ☹ Stop once the model's prediction matches the target label
- ☹ Require a differentiable classifier

**Conditional Generative Model (CGM)-based Methods:**

- ☹ Without control over boundary distance
- ☹ Need fine-tuning whenever the classifier changes

## LeapFactual

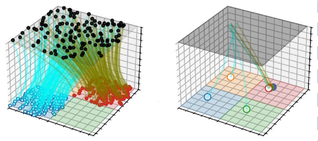




**Training Phase**: We demonstrate that the classifier's output is **disentangled** from the residual by training with our proposed Counterfactual Explanation–Conditional Flow Matching (**CE-CFM**) objective:

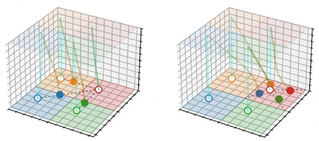
$$\mathcal{L}(\psi) := \mathbb{E}_{t,q(h),p_t(z|h)} \left\| v_\psi(t, z, \boxed{c}) - u_t(z|h) \right\|$$

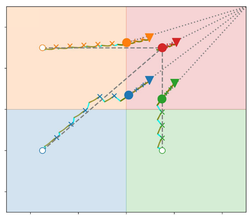
**Explaining Phase:** We propose two algorithms, LEAP and LEAPFACTUAL. Leveraging the flexibility of CFM, LEAP enables information replacement, **blending**, and **injection**. LEAPFACTUAL further integrates blending and injection leaps to generate **reliable** counterfactual explanations.

**Compression** **Replacement**



**Blending** **Injection**





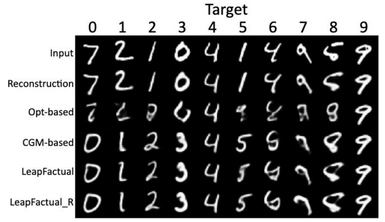
**Algorithm 1** LEAP

Input: Flow matching model $v_\psi$ trained with CE-CFM objective, source point $z_{\text{source}}$ (at $t{=}1$), current label $y_c$, target label $\hat{y}_c$, step sizes $\gamma_{\text{lift}}$, $\gamma_{\text{land}}$
Step 1: Lift ▷ From $Z_1$ to $Z_0$
$z^{y_c}(t) = z_{\text{source}} + \int_1^t \gamma_{\text{lift}} v_\psi(\tau, z^{y_c}(\tau), y_c)\, d\tau, \quad t \in [0,1].$
$z_{\text{lift}} \leftarrow z^{y_c}(0)$
Step 2: Land ▷ From $Z_0$ to $Z_1$
$z^{\hat{y}_c}(t) = z_{\text{lift}} + \int_0^t \gamma_{\text{land}} v_\psi(\tau, z^{\hat{y}_c}(\tau), \hat{y}_c)\, d\tau, \quad t \in [0,1].$
$z_{\text{land}} \leftarrow z^{\hat{y}_c}(1)$
Output: Transported point $z_{\text{land}}$

**Algorithm 2** LEAPFACTUAL

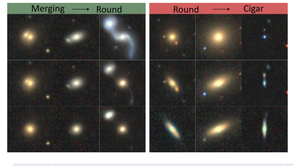
Input: Source point $x$, target label $\hat{y}_c$, classifier $f_\theta$, generative model $g_\phi$, flow matching model $v_\psi$ trained with CE-CFM objective
Hyperparameters: Number of blending leaps $N_{\text{b}}$, blending step size $\gamma_{\text{b}}$, number of injection leaps $N_{\text{i}}$, injection step sizes $\gamma_{\text{i,lift}} < \gamma_{\text{i,land}}$
Step 1: Preprocessing
$z \leftarrow g_\phi^{-1}(x)$ ▷ Acquiring $z$ via VAE encoder or GAN inversion
$y_c \leftarrow f_\theta(g_\phi(z))$ ▷ Determining the current label
Step 2: Information Blending ▷ Generating CE
**for** j = 0 to $N_{\text{b}} - 1$ **do**
$z \leftarrow \text{LEAP}(v_\psi, z, y_c, \hat{y}_c, \gamma_{\text{b}}, \gamma_{\text{b}})$ ▷ Blending the source and target classes information
$y_c \leftarrow f_\theta(g_\phi(z))$
**end for**
Step 3: (Optional) Information Injection ▷ Generating Reliable CE
**for** j = 0 to $N_{\text{i}} - 1$ **do**
$z \leftarrow \text{LEAP}(v_\psi, z, y_c, \hat{y}_c, \gamma_{\text{i,lift}}, \gamma_{\text{i,land}})$ ▷ Injecting the target class information
$y_c \leftarrow f_\theta(g_\phi(z))$
**end for**
Step 4: Postprocessing
$x_{\text{CE}} = g_\phi(z)$
Output: Transported point $x_{\text{CE}}$

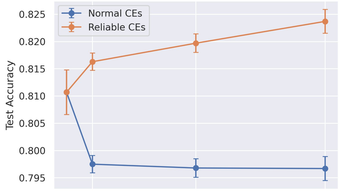
## Experiments




**Quantitative Assessment**: Our methods produce **realistic** counterfactual samples that modify only the most important **class-related features**, preserving digit style. This selective modification is particularly beneficial for distinguishing similar classes (e.g., 9 vs. 7 or 5 vs. 8), thereby **enhancing interpretability**.

**Model Improvement**: We show the advantages of reliable CEs in model training on the Galaxy10 DECaLS. When **standard CEs** are blended into the training data, model performance **declines** as the proportion of CEs increases. In contrast, using **reliable CEs** leads to **improvements** in both accuracy and AUC with increasing fraction. This is because the standard CEs are on the learned decision boundary of the weak classifier, while reliable CEs align closely with the true decision boundary. They can serve as an augmentation strategy for imbalanced classes, addressing fairness.




**Generalization**: We show that the proposed method is highly scalable (by using an SOTA generative model like Style GAN) and applicable to non-differentiable classifiers, such as human annotators (proxied by a pretrained CLIP model)