Introduction to
Counterfactual Explanations

Mateusz Krzyziński
Bartek Sobieski

MI

Warsaw, April 8th, 2024

Agenda

1. Intuition & definition
2. Properties
3. CEs for tabular data
4. CEs for images

Note, much of this presentation is based on the article:
R. Guidotti, ***Counterfactual explanations and how to find them: literature review and benchmarking***,
Data Mining and Knowledge Discovery, 2022.

What are Counterfactual Explanations?

Counterfactual thinking

AMERICAN PSYCHOLOGICAL ASSOCIATION

counterfactual thinking

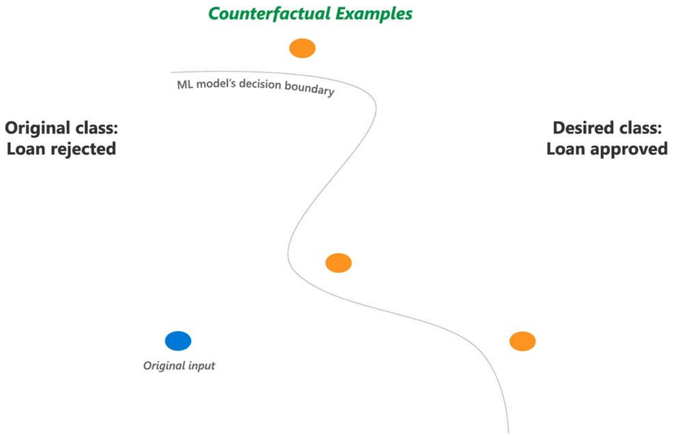
Updated on 04/19/2018

1. imagining ways in which events in one's life might have turned out differently. This often involves feelings of regret or disappointment (e.g., *If only I hadn't been so hasty*) but may also involve a sense of relief, as at a narrow escape (e.g., *If I had been standing three feet to the left…*).

2. any process of reasoning based on a conditional statement of the type "If X, then Y" where X is known to be contrary to fact, impossible, or incapable of empirical verification. Counterfactual thinking of the first sort is common in such historical speculations as *If Hitler had been killed in July 1944, then …* . Counterfactual thinking of the second and third types can play a useful role in evaluating the implications of a theory or heuristic and in thought experiments. See also as-if hypothesis; conditional reasoning.

Counterfactual explanations – intuition

Counterfactual explanations suggest
what should be different in the input instance
to change the prediction of a model.



Amit Sharma, DiCE, Microsoft Research Blog

Counterfactual explanations – formalization

Given a model f that outputs the prediction $y = f(x)$ for an instance x, a counterfactual explanation consists of an instance x' such that the prediction for f on x' is different from y, i.e., $y \neq f(x')$, and such that the difference between x and x' is *minimal*.

- ❑ post-hoc
- ❑ example-based
- ❑ local*
- ❑ both model agnostic and model specific

*sets of CEs and their aggregations can be considered global explanations

We will discuss global approach later in this track.

Counterfactual explanations – first method

Counterfactual instance is to be find based on the optimization:

$$\mathrm{x}_c = \arg\min_{\mathrm{x}'} \{\max_{\lambda} \lambda(\hat{f}(\mathrm{x}') - y')^2 + d(\mathrm{x}, \mathrm{x}')\}$$

parameter to balance the trade-off

desired prediction

distance measure
L1 weighted with the inverse median absolute deviation

There are only(?) two properties considered here.

S. Wachter, B. Mittelstadt, C. Russell, ***Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR***, Harvard Journal of Law & Technology, 2018.

Properties of
Counterfactual Explanations

Desirable properties

Validity.

A counterfactual *x′* is valid iff it actually changes the prediction with respect to the original one, i.e., $f(\mathrm{x}) \neq f(\mathrm{x}')$

Minimality (Sparsity).

x′ is minimal iff there is no other valid example *x″* with the smaller number of different attributes w.r.t. ***x***, i.e., $\forall \mathrm{x}'' \quad |\delta_{\mathrm{x},\mathrm{x}''}| \geq |\delta_{\mathrm{x},\mathrm{x}'}|$

Similarity (Proximity).

x′ should be similar to x, i.e., $d(\mathrm{x}, \mathrm{x}') < \varepsilon$ given a distance function d and a predefined threshold ε

Desirable properties cont'd

Plausibility.

A counterfactual x' is plausible iff it is coherent with a reference sample, i.e., it is not an outlier, out-of-distribution example.

Actionability.

A counterfactual x' is actionable iff all the differences between x' and x are related to actionable features that can be mutated/changed.

Diversity.

If a set of counterfactual examples is returned, it should be formed by diverse examples (similar to x but most different between each other).

Desirable properties cont'd

Discriminative power. (Warning! Highly subjective)

A counterfactual x' must help in figuring out why different prediction can be obtained with it. Difficult to quantify without human-based evaluation.

Causality.

A counterfactual x' should maintain any known causal relationship between features (related to plausibility).

How to retrieve Counterfactual Explanations?

Strategies for finding counterfactuals

Optimization

minimizing a loss function that accounts for desired properties

Heuristic Search

local and heuristic choices performed iteratively

Instance-Based

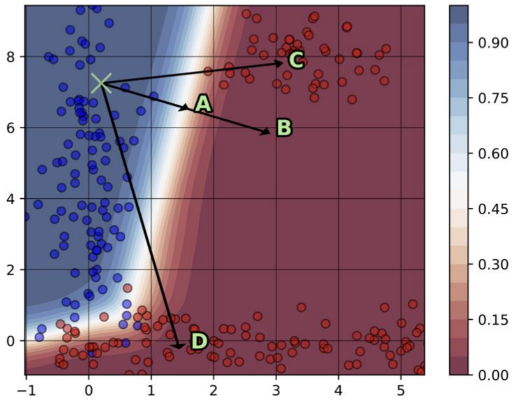
selecting the best example from a reference dataset

Decision Tree

exploiting the structure of a decision tree approximating the black-box

Strategies for finding counterfactuals

Instance-Based



FACE: Feasible and Actionable Counterfactual Explanations

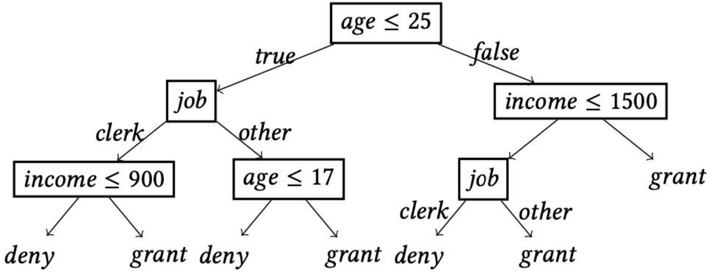
1. Construction of a weighted graph over data points (KDE, kNN, ε-graph).
2. Preparing the list of candidate targets (based on additional constraints).
3. Finding shortest path in a constructed graph for each candidate target.
4. Selecting the candidate with the shortest path.

We will discuss attacks on instance-based CEs next week.

R. Poyiadzi et al., *FACE: Feasible and Actionable Counterfactual Explanations*, AIES, 2020.

Strategies for finding counterfactuals

Decision Tree



LORE: Local Rule-Based Explanations

1. Generate a set of synthetic samples from the neighbourhood of ***x*** through a genetic algorithm.
2. Train a decision tree on this set labeled with the black-box predictions.
3. Extract the counterfactual rules from a trained decision tree (those with the smallest number of split conditions not satisfied by ***x***).

R. Guidotti et al., ***Local Rule-Based Explanations of Black Box Decision Systems***, IEEE Intelligent Systems, 2019.

Counterfactual explanations in computer vision

Definition

Given a **classifier** and an **image**, what is the **minimal semantic change** that **flips the model's decision**.

Prediction: smiling

Prediction: not smiling

Tabular vs visual

Curse of dimensionality

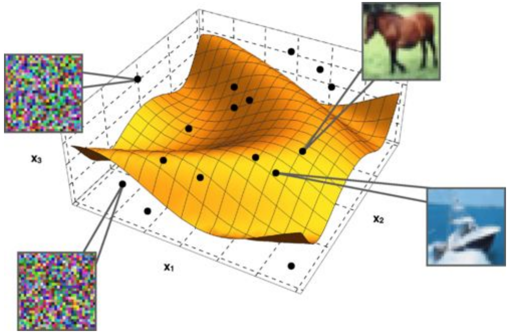
	Tabular	Visual
Dimensionality	~ 10 - 1000 features	~ 50 000 - 200 000 pixels

Visual data is **much more sparse** than tabular data

Manifolds differ

Given an observation of each modality, move along a random direction:

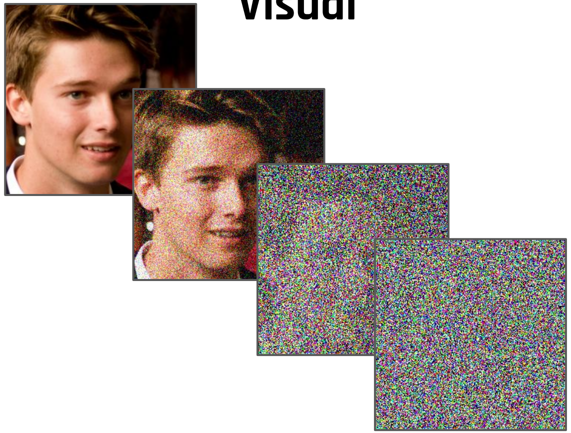
$$\mathbf{x} + \alpha \cdot \mathbf{d}, \text{ where } \|\mathbf{d}\| = 1, \alpha = \beta \cdot n_{\text{features}}$$



Tabular

$\mathbf{X}$	Age: 44	Salary (k): 15	Flat size (m^2): 96	Savings (k): 12
$\beta = 0.001$	Age: 43.99	Salary (k): 14.9994	Flat size (m^2): 95.999	Savings (k): 11.999
$\beta = 0.01$	Age: 43.96	Salary (k): 14.994	Flat size (m^2): 95.99	Savings (k): 11.99
$\beta = 0.1$	Age: 43.65	Salary (k): 14.94	Flat size (m^2): 95.9	Savings (k): 11.88

Visual



$\mathbf{X}$

$\beta = 0.000001$

$\beta = 0.00001$

$\beta = 0.0001$

Modality-specific differences

$$\mathrm{x}_c = \arg\min_{\mathrm{x}'} \{\max_{\lambda} \ \lambda(\hat{f}(\mathrm{x}') - y')^2 + d(\mathrm{x}, \mathrm{x}')\}$$

?

?

How can the differences between visual and tabular data be accounted for in the optimization objective?

Measuring perceptual distance

Semantic similarity

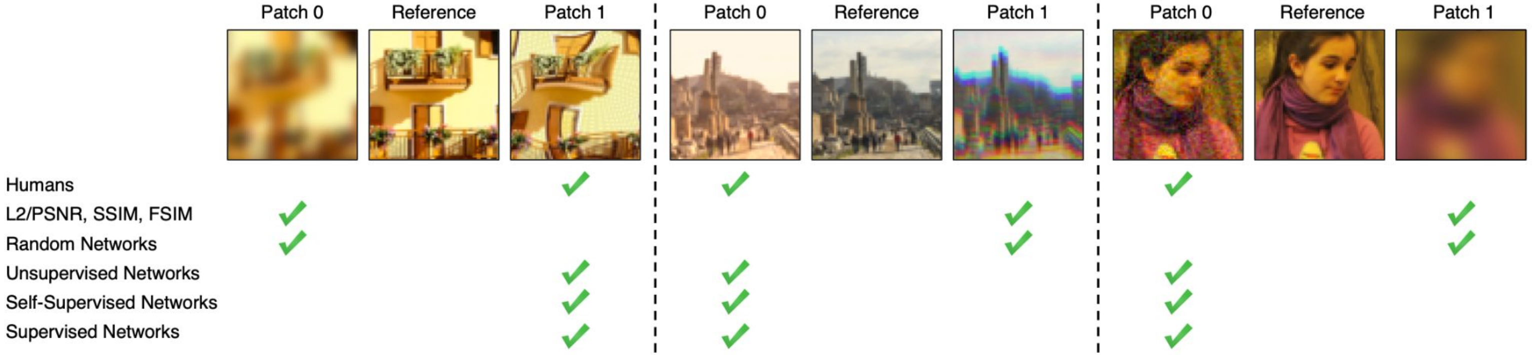
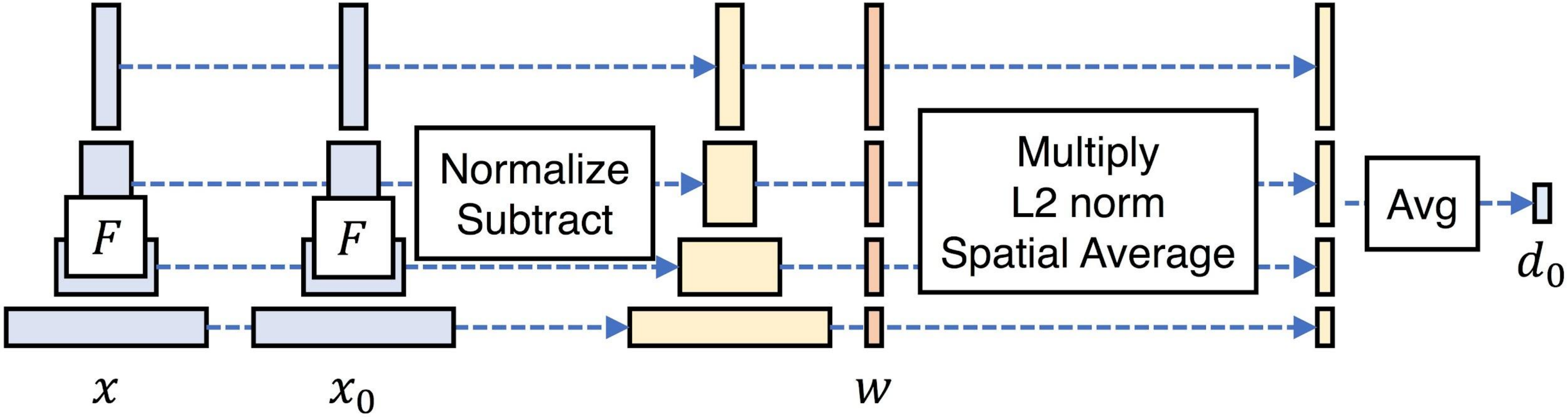


Figure 1: **Which patch (left or right) is "closer" to the middle patch in these examples?** In each case, the traditional metrics (L2/PSNR, SSIM, FSIM) disagree with human judgments. But deep networks, even across architectures (Squeezenet [20], AlexNet [27], VGG [52]) and supervision type (supervised [47], self-supervised [13, 40, 43, 64], and even unsupervised [26]), provide an *emergent embedding* which agrees surprisingly well with humans. We further calibrate existing deep embeddings on a large-scale database of perceptual judgments; models and data can be found at `https://www.github.com/richzhang/PerceptualSimilarity`.

R. Zhang, P. Isola, A. Efros, E., Shechtman, O. Wang, ***The Unreasonable Effectiveness of Deep Features as a Perceptual Metric***, CVPR, 2018.

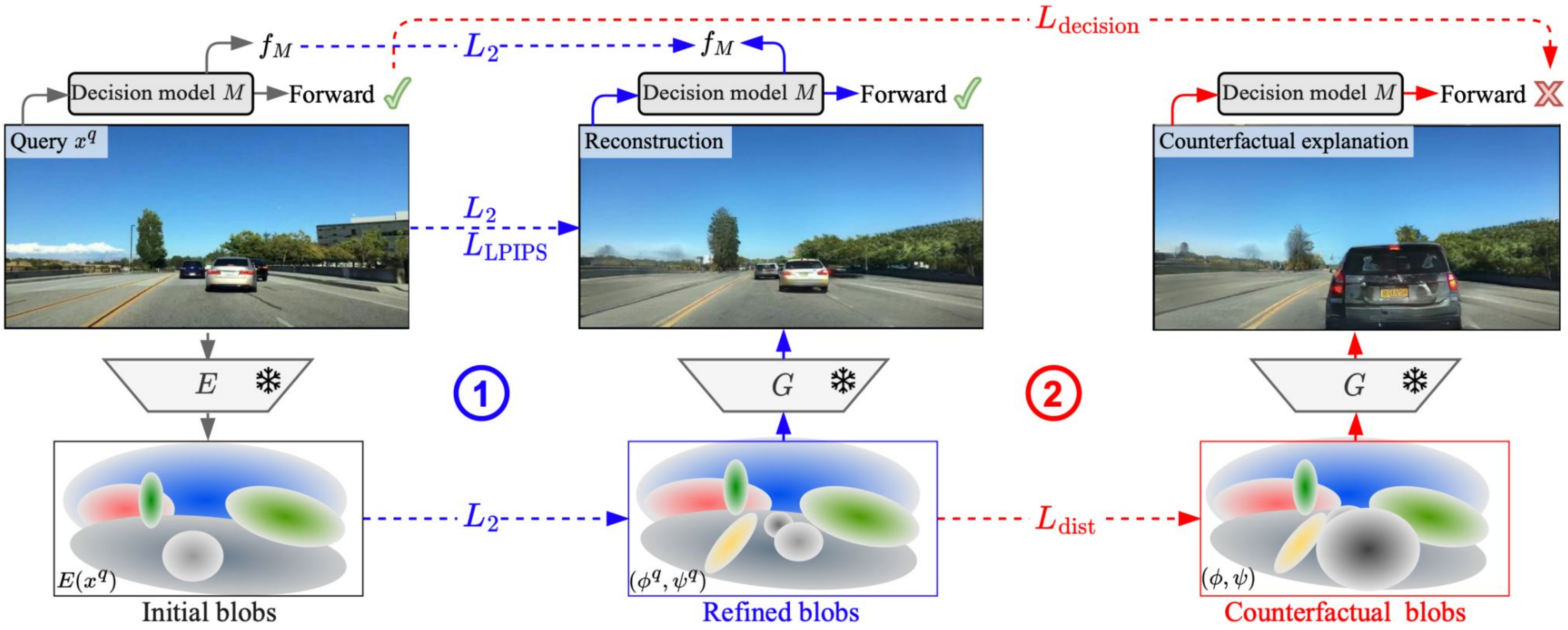
LPIPS



R. Zhang, P. Isola, A. Efros, E., Shechtman, O. Wang, ***The Unreasonable Effectiveness of Deep Features as a Perceptual Metric***, CVPR, 2018.

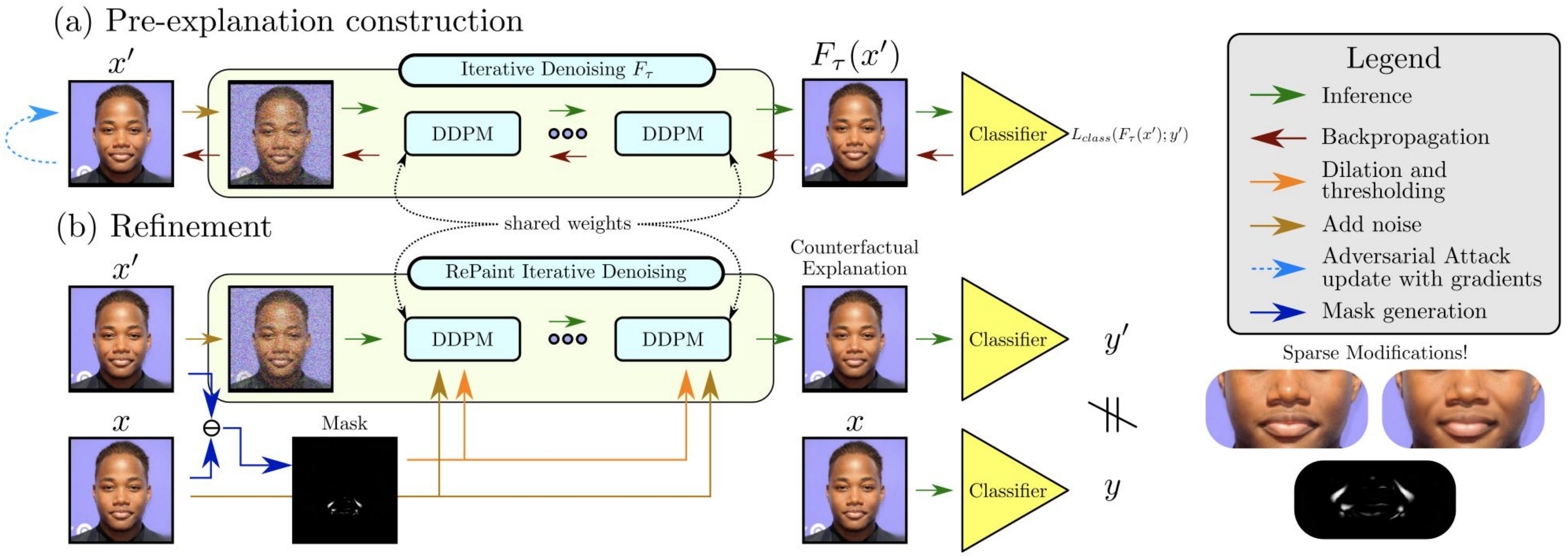
State-of-the-art in generating explanations

OCTET: Object-aware Counterfactual Explanations



M. Zemni, M. Chen, E. Zablocki, H. Ben-Younes, P. Perez, M. Cord, ***OCTET: Object-aware Counterfactual Explanations***, CVPR, 2023.

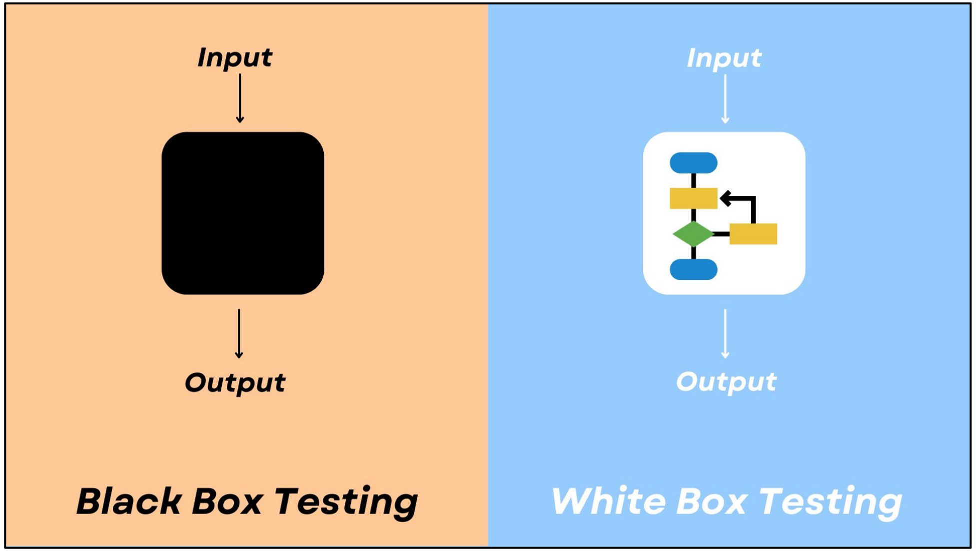
Adversarial Counterfactual Visual Explanations



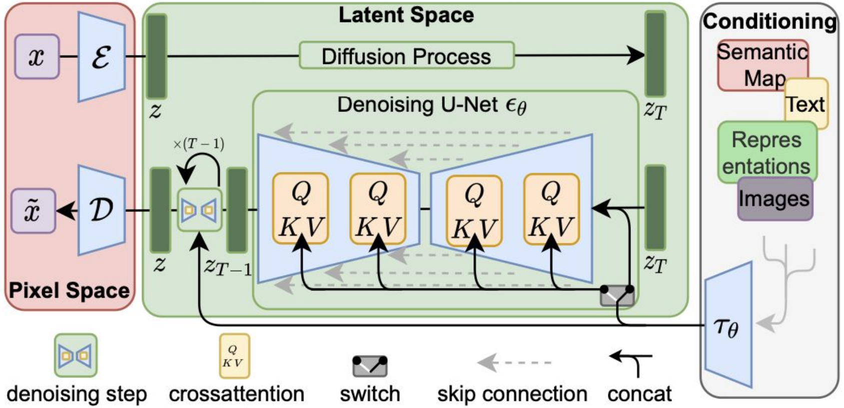
G. Jeanneret, L. Simon, F. Jurie, ***Adversarial Counterfactual Visual Explanations***, CVPR, 2023.

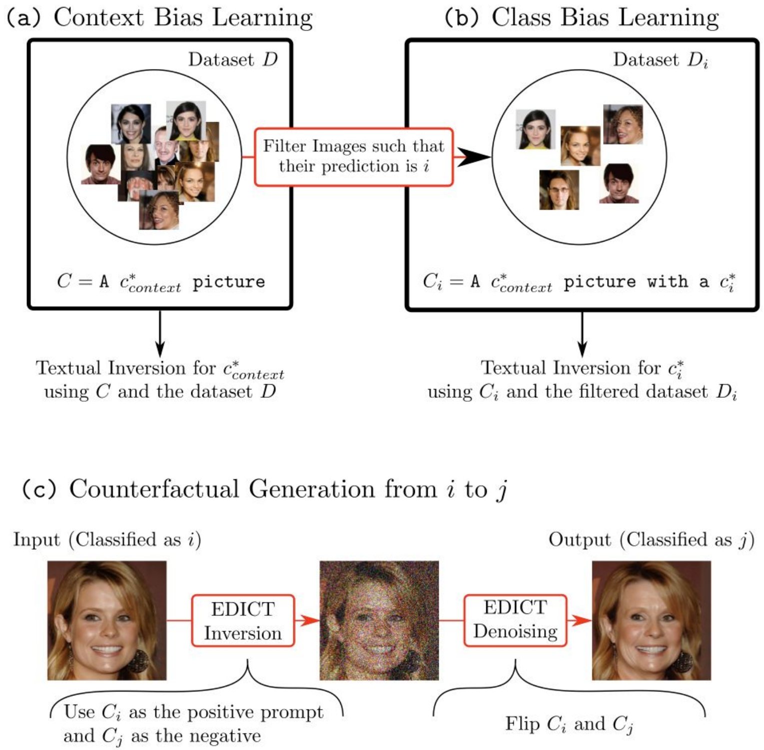
Current challenges

Lack of black-box methods



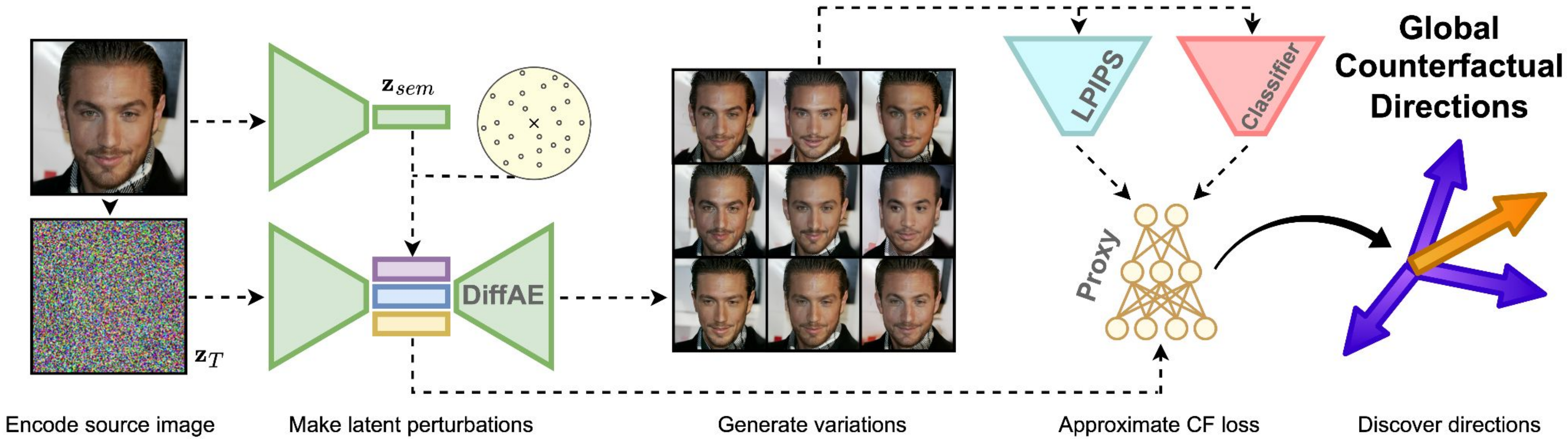
TiME





G. Jeanneret, L. Simon, F. Jurie, ***Text-to-image Models for Counterfactual Explanations: a Black-Box approach***, WACV, 2024.

GCD



$\mathbf{z}_{sem}$
$\mathbf{z}_T$
DiffAE
LPIPS
Classifier
Proxy
Global Counterfactual Directions
Encode source image
Make latent perturbations
Generate variations
Approximate CF loss
Discover directions

Problematic evaluation

There is no perfect metric to evaluate visual counterfactual explanations and many of the existing ones are just noise

Thank you for attention