Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

On generating trustworthy counterfactual explanations

Javier Del Ser [a,b], Alejandro Barredo-Arrieta [c], Natalia Díaz-Rodríguez [d], Francisco Herrera [d], Anna Saranti [e,f], Andreas Holzinger [e,f,g,*]

[a] TECNALIA, Basque Research and Technology Alliance (BRTA), P. Tecnologico, Ed. 700, 48160 Derio, Spain
[b] Department of Communications Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain
[c] Kurago Software, 48011 Bilbao, Spain
[d] DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, 18071 Granada, Spain
[e] Human-Centered AI Lab, University of Natural Resources and Life Sciences Vienna, A-1190 Vienna, Austria
[f] Medical University Graz, A-8036 Graz, Austria
[g] xAI Lab, Alberta Machine Intelligence Institute, Edmonton, Canada

ARTICLE INFO

Keywords:
Explainable artificial intelligence
Deep learning
Counterfactual explanations
Generative adversarial networks
Multi-objective optimization

ABSTRACT

Deep learning models like chatGPT exemplify AI success but necessitate a deeper understanding of trust in critical sectors. Trust can be achieved using counterfactual explanations, which is how humans become familiar with unknown processes; by understanding the hypothetical input circumstances under which the output changes. We argue that the generation of counterfactual explanations requires several aspects of the generated counterfactual instances, not just their counterfactual ability. We present a framework for generating counterfactual explanations that formulate its goal as a multiobjective optimization problem balancing three objectives: plausibility; the intensity of changes; and adversarial power. We use a generative adversarial network to model the distribution of the input, along with a multiobjective counterfactual discovery solver balancing these objectives. We demonstrate the usefulness of six classification tasks with image and 3D data confirming with evidence the existence of a trade-off between the objectives, the consistency of the produced counterfactual explanations with human knowledge, and the capability of the framework to unveil the existence of concept-based biases and misrepresented attributes in the input domain of the audited model. Our pioneering effort shall inspire further work on the generation of plausible counterfactual explanations in real-world scenarios where attribute-/concept-based annotations are available for the domain under analysis.

1. Introduction

In recent years, Deep Neural Networks (commonly referred to as “Deep Learning”) have made significant advancements, surpassing theoretical analysis of their properties and finding implementation and utilization in a diverse range of real-world applications, including agriculture [25], health [2], or life sciences [22]. Currently, the success of Deep Learning is culminating in natural language understanding and response. This has been made possible by a special class of Deep Learning called Transformers. The Transformer model uses a self-observation mechanism that allows it to detect dependencies between different elements within a sequence. This mechanism allows the model to weigh the relevance and importance of different items in the input sequence when making pre-

[*] Corresponding author at: University of Natural Resources and Life Sciences Vienna, Peter Jordan Str. 82, A-1190 Wien, Austria.
E-mail address: andreas.holzinger@human-centered.ai (A. Holzinger).

https://doi.org/10.1016/j.ins.2023.119898
Received 29 May 2023; Received in revised form 9 November 2023; Accepted 10 November 2023

Available online 17 November 2023

dictions. Unlike traditional recurrent neural networks (RNNs) that process sequences sequentially, the Transformer can process all positions in parallel, making it extremely efficient at capturing long-range dependencies. The core idea behind the Transformer is based on the concept of self-attention or scaled dot-product attention. It computes attention weights between each pair of items in the input sequence, allowing the model to pay attention to different parts of the sequence when generating the output. This is the basis for the large language models that are so successful today, such as Generative Pre-trained Transformer (GPT), BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT approach), etc. These successes are used to perform human-like conversations.

Particularly in dealing with high-dimensional data, Deep Learning has demonstrated promising outcomes, revolutionizing the landscape of machine learning modelling approaches. Its superior performance has been observed in numerous scenarios involving image, video, and spatial-temporal data, and has become crucial in the domain of neural graph networks for graph data.

Unfortunately, some concerns arise from the mismatch between research studies dealing with Deep Learning applied to certain modelling tasks (*let the model perform to its best for the task at hand*) and the real-world use of models to improve an already known solution. Most in-field approaches contemplate attempts at improving an already human-created solution to solve a problem (optimizing a process), whereas the most common Deep Learning approaches are better suited to find their own solutions to a more high-level problem (predicting an outcome). Together with this difference, another concern deals with the difficulty of understanding and interpreting the mechanisms by which Deep Learning works, particularly when the audience that makes decisions on their outputs lacks any knowledge about Computer Science and Data Science. This renders Deep Learning a less useful modelling choice for real-world scenarios in which models are used to improve decision-making in processes that are managed by humans and/or where decisions affect humans, as in life sciences, law or social policy-making, among others. In other words, the actionability of predictions requires a step beyond a proven good generalization performance of the model issuing them [40].

In order to bridge this gap in Machine Learning-based decision-making, new frameworks for explainability are required. These frameworks aim to give insights not only to experts in the field of application but also to those commonly in charge of the use and maintenance of the deployed models. These two audience profiles differ significantly in what refers to their capabilities to understand explanations generated for a given model. These different capabilities entail that approaches to explain Deep Learning models generate explanations better suited for auditing the models by developers, leaving them far from the cognitive requirements of experts that ultimately make decisions in practice.

Recent research is profoundly bothered with bridging this gap. To this end, the broad scope of model explainability has been approached from manifold areas, including robustness by adversarial attacks [18], uncertainty estimation [17], visualization of internal representations [42] or attention-based explanations [4]. Even though the research community is thrilled with new advances in explainability, they do not entirely bridge the aforementioned gap between theoretical developments and their practical adoption. Most explainability solutions [5] consider an audience with profound knowledge of the inner workings of the models, which eases the understanding of explanations but does not comply with real-world settings often encountered in model-based decision-making processes.

Among the alternatives to reach such a universal understanding of model explanations, counterfactual examples are arguably the one that best conforms to human understanding principles when faced with unknown phenomena, because there is evidence from human reasoning [46], [9]. The underlying concept in our paper is human counterfactual thinking, which describes a set of possible alternatives to events that have already occurred, but which contradict the actual events [39]. Indeed, discerning what would happen should the initial conditions differ in a plausible fashion is a mechanism often adopted by a human when attempting to understand any unknown phenomenon [45]. Circumscribing the factual boundaries by which a given model works *as usual* can be conceived as a post-hoc explainability method, which is grounds on an adversarial analysis of the audited model [43]. From the practical perspective, several aspects of the produced counterfactual examples should be considered besides their plausibility, so that the audience of the model can examine the limits of the model from a multi-faceted perspective. It is only by investigating this manifold interplay between the features of the generated counterfactual explanations that a well-rounded counterfactual analysis can be achieved.

This manuscript joins the research area aimed at making Deep Learning models more usable in practice via counterfactual-based explanations. To this end, we propose an adversarial strategy to produce counterfactual examples for a Deep Learning classifier. This classifier to be audited solves a task defined over a certain dataset (e.g. discriminating male and female images from human faces) so that counterfactual explanations are generated to explain the boundaries of the model once trained to address the classification task at hand. We further impose that the generated counterfactual examples are *plausible*, i.e., changes made on the input to the classification model have an appearance of credibility to humans without any computer intervention. To ensure plausibility, the proposed method makes use of GANs (Generative Adversarial Networks) in order to learn the underlying probability distribution of each of the features needed to create examples of a target distribution (namely, human faces). Our framework allows searching among samples of the first distribution to find realistic counterfactual explanations close to a given test sample that could be misclassified by the model (namely, the face of a male being classified as a female). As a result, our framework makes the user of the model assess its limits with an adversarial analysis of the probability distribution learned by the model, yet maintaining a sufficient level of plausibility for the analysis to be understood by a non-expert user. As a step beyond the state of the art, the proposed framework ensures the production of multi-faceted counterfactual examples by accounting for two additional objectives besides plausibility: the *intensity of the modification* made to an original example to produce its counterfactual version; and its *adversarial power*, which stands for the change in the output of the model that is audited.

In summary, the main contributions of this work can be summarized as follows:

Table 1
Summary of symbols, meaning and their first appearance in the manuscript.

Symbol	Appearance	Meaning
$\mathbf{x}$	Section 3.2	Input example to the audited model
$P_{\mathbf{X}}(\mathbf{x})$	Section 3.2	Input data distribution followed by $\mathbf{x}$
$\mathbf{a}$	Section 3.2	Original attribute vector
$\mathbf{b}$	Section 3.2	Modified attribute vector
$\boldsymbol{\delta}$	Section 3.2	Perturbation vector ($\boldsymbol{\delta} = \mathbf{b} - \mathbf{a}$)
$\mathbf{x}^{\mathbf{b},\prime}$	Section 3.2	Generated counterfactual for input $\mathbf{x}^{\mathbf{a}}$ and attribute vector $\mathbf{b}$
$\widehat{\mathbf{b}}$	Section 3.2	Modified attribute vector predicted by $C(\cdot)$
$\oplus$	Section 3.2	Superscript defining the anchor sample for which a counterfactual is produced
$C(\cdot)$	Section 3.2, Fig. 2	Classifier that predicts the attribute vector of its input query
$D(\cdot)$	Section 3.2, Fig. 2	Discriminator module of a GAN
$G_{enc}(\cdot)$	Section 3.2, Fig. 2	Encoder of a GAN generator module
$G_{dec}(\cdot)$	Section 3.2, Fig. 2	Decoder of a GAN generator module
$T(\cdot)$	Section 3.2, Fig. 2	Target Model Under Test (MUT)
$\mathcal{L}_{rec}(\mathbf{x},\mathbf{x}^{\mathbf{a}\prime})$	Eqs. (1) and (2)	Reconstruction loss
$\mathcal{L}_{att}^{G}(\mathbf{b},\widehat{\mathbf{b}}')$	Eqs. (1) and (3)	Attribute loss
$\mathcal{L}_{adv}^{G}(\mathbf{x}^{\mathbf{b}\prime})$	Eqs. (1) and (4)	Adversarial loss
λ_i	Eqs. (1) and (5)	Weights of reconstruction ($i = 1$) and attribute terms ($i = 2, 3$) in the training losses of $G_{enc}(\cdot)$, $G_{dec}(\cdot)$, $D(\cdot)$ and $C(\cdot)$
$f_{att}(\cdot)$	Eq. (11)	Function quantifying the *change intensity* of the generated counterfactual
$f_{gan}(\cdot)$	Eq. (11)	Function quantifying the *plausibility* of the generated counterfactual
$f_{adv}(\cdot)$	Eq. (11)	Function quantifying the *adversarial power* of the generated counterfactual

- We present a novel framework to generate multi-faceted counterfactual explanations for targeted classification models. The framework brings together GAN architectures for generative data modelling and multi-objective optimization for properly balancing among possible conflicting properties sought for the counterfactuals: plausibility, change intensity and adversarial power.
- The framework is described mathematically, and the design rationale for each of its compounding blocks is given and justified.
- Explanations generated by the framework are showcased for several classifiers and GAN models for image and volumetric data, answering with empirical evidence three research questions: Q1) If there effectively exists a trade-off between the properties of the counterfactual set; Q2) whether the counterfactual explanations align well with human logic that can be articulated around the input domain of the audited model; and Q3) if the produced counterfactual explanations can be of any help beyond explainability itself.
- Throughout the six experiments discussed in the manuscript we argue and show that, when inspected from a multi-faceted perspective, counterfactual explanations can be an informative human-centred tool for concept-defined bias analysis and the discovery of misrepresentations in the data space.

The remainder of the article is organized as follows: Section 2 covers some background required for connecting the four core aspects of our proposed framework: Deep Learning for image classification, GANs, model explainability and counterfactual explanations. Section 3 details the framework proposed in this study, including a mathematical statement of the problem tackled via multi-objective optimization and a discussion on how the output of the framework can be consumed by different audiences. Section 4 describes the experimental setup designed to showcase the output of the framework. Section 5 presents and discusses the results stemming from the performed experiments. Finally Section 6 draws conclusions and future research lines related to our findings. Table 1 summarizes the main mathematical symbols used throughout the manuscript.

2. Background

As anticipated in the introduction, the proposed framework resorts to GANs for producing realistic counterfactual examples of classification models. Since the ultimate goal is to favour the understanding of the model classification boundaries by an average user, the framework falls within the XAI (Explainable Artificial Intelligence) umbrella. This section briefly contextualizes and revisits the state-of-the-art research areas related to the framework: Deep Learning for image classification and generative modelling (Subsection 2.1), XAI and counterfactual analysis (Subsection 2.2) and multi-objective optimization (Subsection 2.3).

2.1. Deep learning for image classification and generative modelling

When it comes to classification tasks over image data, the reportedly superior modelling capabilities of Convolutional Neural Networks (CNNs) are often adopted to capture spatial correlations in image data by learning hierarchical filters. This is achieved by virtue of trainable convolutional filters which can be trained via gradient backpropagation or even imported from other networks pretrained for similar tasks, giving rise to image classification models of the highest performance. The increasing availability of image datasets and the capability of processing them efficiently have yielded hierarchically stacked CNNs that, despite attaining unprecedented levels of performance, come at the cost of more complex, less understandable model structures [31]. The more complex the model is, the harder is to pinpoint the reasons for its failures. The need for auditing these black boxes is the core motivation of the study presented in this paper.

Another task for which CNNs are crucial is generative modelling, e.g. the construction of models capable of characterizing the distribution of a given dataset and sampling it to create new, synthetic data instances. When the dataset is composed of images, generative adversarial networks (GANs) are arguably the spearhead in image generative modelling. GANs were first introduced by Goodfellow [18], bringing the possibility of using neural networks (*function approximators*) to become generators of a desired distribution. Since their inception, GANs have progressively achieved photo-realistic levels of resolution and quality when synthesizing images of different kinds. In general, a GAN architecture consists of two data-based models, which are trained in a mini-max game: one of the players (models) minimizes its error (loss), whereas the other maximizes its gain. In such a setup, multiple models have flourished to date, each governed by its strengths and vulnerabilities. In connection to the scope of this paper, some of these were conceived with the intention of finding the adversarial boundaries of a certain model. Other GAN approaches aim at generating samples of incredibly complex distributions like photo-realistic human faces.

As will be later detailed in Section 3, the framework proposed in this work hybridizes these two uses of CNNs by optimizing the output generated by a GAN to perform a counterfactual analysis of a given classification model to be audited.

2.2. Explainable artificial intelligence (XAI) and counterfactual explanations

Model explainability [5] and causability [23] have recently become topics of capital importance in Machine Learning, giving rise to a plethora of different approaches aimed at explaining how decisions are issued by a given model. Most research activity in this area is arguably focused on post-hoc XAI tools that produce explanations for single data instances (what is referred to as *local explanations*). The LIME tool is one of this kind, visualizing a model's internal activations when processing a given test sample. A similar approach is followed by LRP (Layer-wise Relevance Propagation) embedded in the SHAP suite, which highlights the parts of an input image that push the output of the model towards one label or another. This provides an understandable interface of the reasons why the model produces its decision. More recently, Grad-CAM and its successor Grad-CAM++ can be considered as the *de facto* standard for the explainability of local decisions, particularly in the field of image classification. These two methods implement a gradient-based inspection of the knowledge captured by a neural network, giving rise to a quantitative measure of the importance of parts of the image for the output of the model. Unfortunately, the dependence of such explanations on the gradient of the model restricts the applicability of these techniques to other techniques beyond neural architectures.

When pursuing model-agnostic local explanations, a common strategy is to analyze the model from a counterfactual perspective. Counterfactual exploration is an innate process for the human being when facing an unknown phenomenon, system or process [36]. The concept behind counterfactual explanations reduces to providing an informed answer to a simple question: *which changes would make the output of the unknown model for a given input vary?*. Such changes constitute a counterfactual example, always related to an input to the process or system under focus. Based on this concept, many contributions have hitherto developed different XAI approaches to generate counterfactual examples that allow an understanding of how Machine Learning models behave. Some approaches are based on discovering the ability of a given individual to change the model's outcome. One example is the work in [47], which presents a simple but effective distance-based counterfactual generation approach, that can be used to audit different classifiers (e.g. neural networks and support vector machines). Later, the counterfactual problem is tackled in [44] departing from the premise that a user should be able to change a model's outcome by actionable variables. This hypothesis is validated over linear classifiers but also claimed to be extensible to non-linear classifiers by means of local approximations. In a similar fashion, [27,37] allows the user to guide the generation of counterfactual examples by imposing forbidden changes that cannot be performed along the process. A subset of counterfactual studies is rather focused on the problem of predictive multiplicity [7,33]. Multiple classifiers may output the same solution while treating the data in different ways, hence the generation of counterfactuals can lead to insights into the question of which of these classifiers is better for the problem at hand. In this research area several contributions [35,38,12,29] have developed different schemes to address this problem. Connectedness, proximity, plausibility, stability and robustness as well as other Explainable AI (xAI) metrics [41] are yet other concerns that have pushed the development of techniques for the generation of counterfactuals. In their search for robust interpretability, the work in [3] came up with a method to generate self-explaining models based on explicitness, faithfulness and stability.

Following the extensive analysis carried out in [13], it is of utmost importance to recall the *"master theoretical algorithm"* [19], from which nineteen other algorithms concerning counterfactual explanations can be derived. The nineteen algorithms fall into a categorization of six different counterfactual generation strategies: instance-based, constraint-based, genetic-based, regression-based, game theory-based and case reasoning-based. Instance-based approaches are based on feature perturbance measured by a distance metric. The pitfall of these approaches (when pure) resides in their inability to validate instance plausibility. Constraint-based approaches are, in turn, the methods that modulate their counterfactual search by means of a constraint satisfaction problem. The more general scope of these approaches allows for an easier adaptation to the problems at hand. Genetic-based approaches, as the name conveys, guide the search for counterfactuals as a genetic-oriented optimization problem. Regression-based approaches use the weights of a regression model as a proxy to produce counterfactual examples. However, these approaches again fall short of assuring the plausibility and diversity of the produced counterfactual instances. Game-theory-based approaches are driven by game-theoretical principles (e.g. Shapley values) but also disregard important properties of its counterfactual outputs. Finally, case reasoning-based approaches seek past solutions (in the model) that are close to a given instance, and adapt them to produce the counterfactual. Once again, such adaptations may produce counterfactual instances that, even if close to a certain input, cannot be claimed to be plausible nor diverse with respect to the input under consideration.

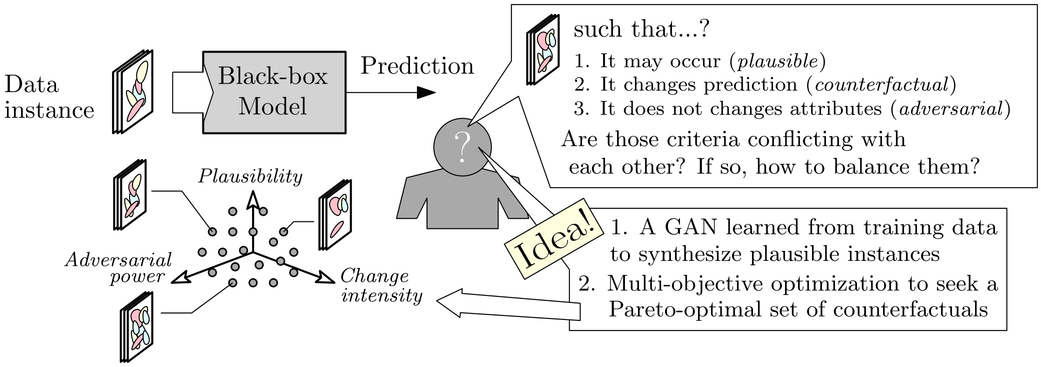

Data instance
Black-box Model
Prediction
such that...?
1. It may occur (plausible)
2. It changes prediction (counterfactual)
3. It does not changes attributes (adversarial)
Are those criteria conflicting with each other? If so, how to balance them?
?
Idea!
1. A GAN learned from training data to synthesize plausible instances
2. Multi-objective optimization to seek a Pareto-optimal set of counterfactuals
Plausibility
Adversarial power
Change intensity

Fig. 1. Conceptual representation of the rationale behind the confluence of predictive modelling, generative adversarial learning, explainability and multi/objective optimization that lies at the core of the proposed framework.

2.3. Multi-objective optimization

From the previous section, it can be inferred that the generation of counterfactual explanations can be mathematically stated as a multi-objective optimization problem comprising different objectives that can conflict with each other. Plausibility – i.e., the likelihood of the counterfactual example to occur in practice – can be thought of as conflicting with the amplitude of the modifications made to the input of the model. Likewise, intense changes in the output of the audited model (namely, its *adversarial power* as introduced in Section 1) when fed with the counterfactual example can jeopardize its plausibility. There lies the contribution of the framework proposed in this work: the generation of a portfolio of counterfactual examples to a certain input that optimally trades among these objectives. This portfolio provides richer information for the user to understand the behaviour of the audited model and distinguishes this work from the current research on counterfactual analysis. The conceptual diagram shown in Fig. 1 illustrates this motivational idea.

To this end, the framework presented in this work falls between constraint-based, genetic-based and instance-based counterfactual explanations, combining these three categories to render a set of multi-criteria counterfactuals. The usage of a GAN architecture presents the ability of a bounded search within a target distribution, enabling quantitative measures of the plausibility of the generated counterfactual (via the discriminator) and algorithmic means to sample this distribution (via the generator). The usage of a multi-objective optimization algorithm yields the ability to guide the counterfactual generation process as per the desired objectives (plausibility, intensity of the modifications and adversarial power), giving rise to the aforementioned portfolio of multi-criteria counterfactual explanations. Among them, we will resort to multi-objective evolutionary algorithms [14], which efficiently perform the search for Pareto front approximations of optimization problems comprising multiple objectives, without requiring information about their derivatives whatsoever.

3. Proposed multi-criteria counterfactual generation framework

This section covers the proposed framework, including the intuition behind its conceptual design (Subsection 3.1), a detailed description of its constituent parts and mathematical components underneath (Subsection 3.2), and an outline of the target audiences that can consume the produced counterfactual explanations, supported by hypothetical use cases illustrating this process (Subsection 3.3). The section concludes with an analysis of the limitations of the proposed framework (Subsection 3.4).

3.1. Design rationale

The explainability framework explores the weaknesses of a target model by means of counterfactual instances generated by a GAN architecture. One of the key aspects of this framework is that it focuses on discovering the reality-bound weaknesses of the target model in the form of examples that, without exiting the realm of plausibility, are able to confound the target model. For instance, for a classifier mapping human faces to their gender (`male`, `female`), the framework can generate modifications of a given input face that are still considered to be real, but they make the audited model change their predicted gender. The overarching motivation of the framework comes from the human inability to assess the working boundaries of a given model in highly-dimensional spaces. In such complex areas, such as image classification, the domain in which images are bound is complex to be characterized, thereby requiring complex generative modelling approaches capable of modelling it and drawing new samples therefrom. The generator of a GAN architecture serves this purpose, whereas the discriminator of the GAN allows verifying whether an output produced by the generator is close to the distribution of the dataset at hand, hence giving an idea of the plausibility of the generated instance.

At this point, it is worth pausing at the further insights that the GAN-based framework can provide. Modifications of an input image producing a counterfactual can be edited by changing the value of variables that affect the output of the GAN generator. Such variables can represent attributes of the input image that ease the interpretation of the results of the counterfactual study regarding the existence of misrepresentations of the reality captured in the dataset at hand and transferred to the audited models. For instance, in the face-gender classifier exemplified previously, let us consider a GAN model with editable attributes (e.g. an AttGAN [21]), including colour hair, face colour or facial expressions. A counterfactual study of a `man` face could reveal that for the face to be classified as a `woman`, the colour hair attribute of all produced counterfactuals shares the same value (*blonde*). Besides the inherent interpretative value of the counterfactuals themselves, our framework can also identify data biases that may have propagated and influenced the generalization capabilities of the audited model.

3.2. Structure and modules

Following the diagram shown in Fig. 2, the design of the proposed framework can be split into four main blocks:

- Target model $T(\cdot)$ to be audited, i.e., the classification model for which the counterfactual study is performed.
- A GAN architecture whose generator module allows inducing conditional perturbations on an input data instance $\mathbf{x}^{\mathbf{a},\oplus}$ (anchor) based on an attribute vector $\mathbf{b}$. Its discriminator module $D(\cdot)$ permits to evaluation of the plausibility of a synthetically generated instance.
- An attribute classifier $C(\cdot)$ that predicts the present attributes in any image fed at its input. The attribute classifier is needed for training the conditional generator module of the GAN so that the generation of the instances allows inserting attribute-based modifications into an original instance without compromising the plausibility of the produced counterfactual.
- A multi-objective optimization algorithm that evolves the perturbation vector $\boldsymbol{\delta} = \mathbf{b} - \mathbf{a}$ to be imprinted on the anchor image $\mathbf{x}^{\mathbf{a},\oplus}$ to the best balance between plausibility, adversarial power and change intensity.

The audited model is fed with the counterfactual example produced by the generator model of the GAN architecture, hence its only prerequisite is that the input of the audited model and the output of the generator are of the same dimensions. In what follows we will assume that the target model to be audited is a CNN used for image classification. Nevertheless, the framework can be adapted to audit other models and tasks whenever the output of the GAN discriminator and the input of the audited model are equally sized, and the measure of adversarial power accounts for the change induced by the counterfactual in the prediction of the model.

The GAN is part of the framework in charge of generating the counterfactual explanations fed to the audited model. Therefore, two requirements are set in this module: 1) the discriminator must be trained for a similar data distribution to that of the audited model, and 2) the generator model must be able to generate samples of such a distribution as per an *attribute vector* $\mathbf{b}$ that controls specific features of the generated instance (image). This attribute vector is tuned by the multi-objective optimization algorithm seeking to maintain plausibility as per the discriminator, changing the output of the audited model and minimizing the intensity of the changes inserted in the original input image.

At this point, it is important to emphasize that the audited model is left aside from the overall training process of the GAN for several reasons. To begin with, for practicality we assume minimum access to the audited model (black-box analysis). Therefore, the logits of the audited model are exploited with no further information on its inner structure. Furthermore, the goal of the discriminator is to decide whether the generated image follows the distribution of the training set, which must be regarded as a plausibility check. The task undertaken by the audited model can be of different types, for instance, to discriminate among `male` and `female`, `old` and `young` or any other task.

The above three-objective optimization problem can be formulated mathematically as follows: let us denote an image on which the counterfactual analysis is to be made as $\mathbf{x}^{\mathbf{a}} \sim P_{\mathbf{X}}(\mathbf{x})$, which follows a distribution $P_{\mathbf{X}}(\mathbf{x})$ and has an attribute vector $\mathbf{a} \in \mathbb{R}^N$. The generator of the GAN model is denoted as $G(\mathbf{x}^{\mathbf{a}}, \mathbf{b})$, whose inputs are the actual image $\mathbf{x}^{\mathbf{a}}$ and a desired attribute vector $\mathbf{b}$. In conditional generative models, the generator is generally composed of an encoder G_{enc} and a decoder G_{dec}. However, for some architectures, the model directly departs from a decoder, given the assumption that the latent code is sampled from a known distribution.

Leaving the special cases aside for the sake of a clearer explanation, the image conditionally output by the generator is given by $\mathbf{x}^{\mathbf{b}\prime} = G_{dec}(G_{enc}(\mathbf{x}^{\mathbf{a}}), \mathbf{b})$. Ideally, when fed with the original attribute vector as the target, $\mathbf{x}^{\mathbf{a}\prime} \approx \mathbf{x}^{\mathbf{a}}$, i.e. the reconstructed image $\mathbf{x}^{\mathbf{a}\prime} = G_{dec}(G_{enc}(\mathbf{x}^{\mathbf{a}}), \mathbf{a})$ should resemble $\mathbf{x}^{\mathbf{a}}$ itself. For non-conditional generative architectures, the generated image is given by $\mathbf{x}' = G_{dec}(G_{enc}(\mathbf{x}))$, where the objective is to have $\mathbf{x}' \approx \mathbf{x}$. A discriminator $D(\mathbf{x}^{\mathbf{b},\prime})$ along with a classifier $C(\mathbf{x}^{\mathbf{b},\prime})$ is placed next along the pipeline to determine 1) whether the synthesized image $\mathbf{x}^{\mathbf{b},\prime}$ is visually realistic; and 2) whether the predicted attributes match the input ones. Again, for non-conditional GAN architectures, only the discriminator $D(\mathbf{x}')$ is necessary.

The overall loss function that drives the learning algorithm of the generator and discriminator is defined as a linear combination of the reconstruction and Wasserstein GAN losses. Assuming an encoder-decoder based generator architecture with latent vector z, the training loss for encoder $G_{enc}(\mathbf{x}^{\mathbf{a}})$ and decoder $G_{dec}(\mathbf{z}, \mathbf{b})$ are given by:

$$\min_{G_{enc}, G_{dec}} \lambda_1 \mathcal{L}_{rec}(\mathbf{x}^{\mathbf{a}}, \mathbf{x}^{\mathbf{a}\prime}) + \lambda_2 \mathcal{L}_{att}^{G}(\mathbf{b}, \widehat{\mathbf{b}}') + \mathcal{L}_{adv}^{G}(\mathbf{x}^{\mathbf{b}\prime}), \tag{1}$$

where:

$$\mathcal{L}_{rec}(\mathbf{x}^{\mathbf{a}}, \mathbf{x}^{\mathbf{a}\prime}) = \mathbb{E}_{\mathbf{x}^{\mathbf{a}} \sim P_{\mathbf{X}}(\mathbf{x})} \left[||\mathbf{x}^{\mathbf{a}} - \mathbf{x}^{\mathbf{b}\prime}||_1 \right] \text{ (reconstruction loss)}, \tag{2}$$

$$\mathcal{L}_{att}^{G}(\mathbf{b}, \widehat{\mathbf{b}}') = \mathbb{E}_{\mathbf{x}^{\mathbf{a}} \sim P_{\mathbf{X}}(\mathbf{x}), \mathbf{b} \sim P_{\mathbf{B}}(\mathbf{b})} \left[\sum_{n=1}^{N=|\mathbf{b}|} H(b_n, \widehat{b_n}') \right] \text{ (attribute loss)}, \tag{3}$$

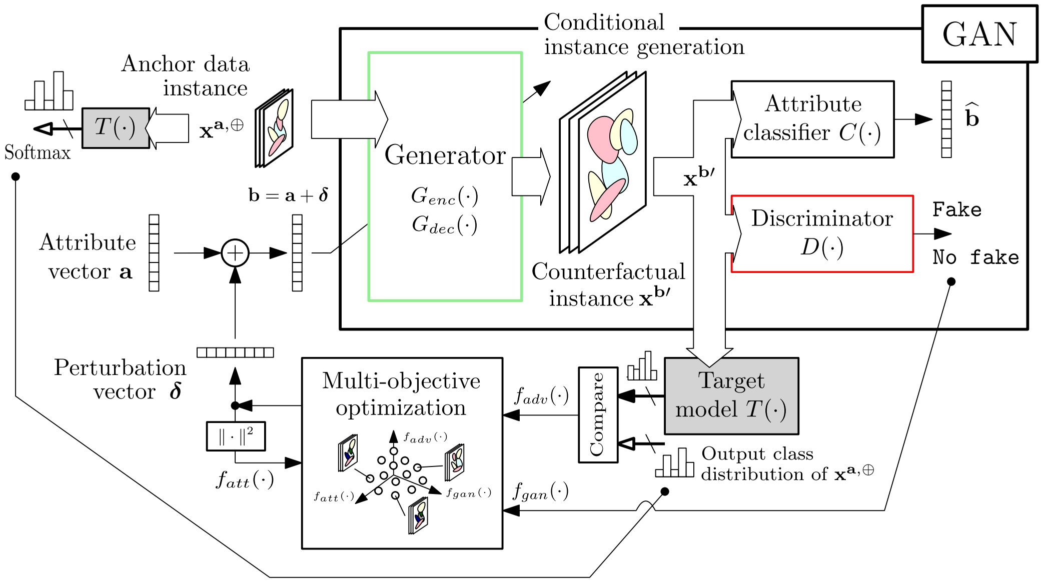

GAN
Conditional instance generation
Anchor data instance
$T(\cdot)$
Softmax
$\mathbf{x}^{\mathbf{a},\oplus}$
Generator
$G_{enc}(\cdot)$
$G_{dec}(\cdot)$
$\mathbf{b}=\mathbf{a}+\delta$
Attribute vector $\mathbf{a}$
Counterfactual instance $\mathbf{x}^{\mathbf{b}'}$
$\mathbf{x}^{\mathbf{b}'}$
Attribute classifier $C(\cdot)$
$\widehat{\mathbf{b}}$
Discriminator $D(\cdot)$
Fake
No fake
Perturbation vector δ
$\|\cdot\|^2$
$f_{att}(\cdot)$
Multi-objective optimization
$f_{adv}(\cdot)$
$f_{att}(\cdot)$
$f_{gan}(\cdot)$
$f_{adv}(\cdot)$
$f_{gan}(\cdot)$
Compare
Target model $T(\cdot)$
Output class distribution of $\mathbf{x}^{\mathbf{a},\oplus}$

Fig. 2. Block diagram of the proposed framework, which is capable of producing counterfactual instances for an audited model $T(\cdot)$ based on three criteria: plausibility, adversarial power and change intensity. Given an anchor data instance and generator, discriminator and an attribute classifier trained over the dataset of the task at hand, the multi-objective solver seeks the set of Pareto-optimal perturbation vectors δ that best balance between the three aforementioned objectives, which are quantified by the already trained blocks of the GAN architecture.

$$\mathcal{L}_{adv}^{G}(\mathbf{x}^{\mathbf{b}'}) = -\mathbb{E}_{\mathbf{x}^{\mathbf{a}}\sim P_{\mathbf{X}}(\mathbf{x}),P_{\mathbf{B}}(\mathbf{b})}\left[D(\mathbf{x}^{\mathbf{b}'})\right] \text{ (adversarial loss).} \tag{4}$$

In the above expressions, $\mathbb{E}[\cdot]$ denotes expectation; $P_{\mathbf{B}}(\mathbf{b})$ indicates the distribution of possible attribute vectors $\mathbf{b}=\{b_n\}_{n=1}^{N}\in\mathbb{R}^N[0,1]$; $H(b_n,\widehat{b}_n{}')$ is the cross-entropy of binary distributions given by b_n and $\widehat{b}_n'\in\widehat{\mathbf{b}}'=C(\mathbf{x}^{\mathbf{b}'})$; and $D(\mathbf{x}^{\mathbf{b}'})=0$ if $\mathbf{x}^{\mathbf{b}'}$ is predicted to be fake.

When it comes to the discriminator $D(\cdot)$ and the classifier $C(\cdot)$, their training loss is given by:

$$\min_{D,C} \lambda_3\mathcal{L}_{att}^{C}(\mathbf{x}^{\mathbf{a}},\mathbf{a}) + \mathcal{L}_{adv}^{D}(\mathbf{x}^{\mathbf{a}},\mathbf{b}), \tag{5}$$

with:

$$\mathcal{L}_{att}^{C}(\mathbf{x}^{\mathbf{a}},\mathbf{a}) = \mathbb{E}_{\mathbf{x}^{\mathbf{a}}\sim P_{\mathbf{X}}(\mathbf{x})}\left[\sum_{n=1}^{|\mathbf{a}|} H(a_n,\widehat{a}_n{}')\right], \tag{6}$$

$$\mathcal{L}_{adv}^{D}(\mathbf{x}^{\mathbf{a}},\mathbf{b}) = -\mathbb{E}_{\mathbf{x}^{\mathbf{a}}\sim P_{\mathbf{X}}(\mathbf{x})}\left[D(\mathbf{x}^{\mathbf{a}})\right] + \mathbb{E}_{\mathbf{x}^{\mathbf{a}}\sim P_{\mathbf{X}}(\mathbf{x}),P_{\mathbf{B}}(\mathbf{b})}\left[D(\mathbf{x}^{\mathbf{b}'})\right], \tag{7}$$

where $\widehat{a}_n{}'\in C(\mathbf{x}^{\mathbf{a}})$, and coefficients $\{\lambda_i\}_{i=1}^{3}$ permit to balance the importance of the above terms during the training process of the GAN architecture. For more general approaches, such as non-conditional GANs, the training loss is given by:

$$\min_{G_{enc},G_{dec}} \lambda_1\mathcal{L}_{rec}(\mathbf{x},\mathbf{x}') + \mathcal{L}_{adv}^{G}(\mathbf{x}'), \tag{8}$$

where:

$$\mathcal{L}_{rec}(\mathbf{x},\mathbf{x}') = \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}(\mathbf{x})}\left[||\mathbf{x}-\mathbf{x}'||_1\right], \tag{9}$$

$$\mathcal{L}_{adv}^{G}(\mathbf{x}') = -\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}(\mathbf{x})}\left[D(\mathbf{x}')\right], \tag{10}$$

and again, coefficient $\lambda_1\in\mathbb{R}[0,1]$ allows tuning the relative importance of the reconstruction loss when compared to the adversarial loss. Once these losses have been defined, the GAN is trained via back-propagation to minimize the losses in Expressions (1) and (5) when measured over a training dataset.

Once trained, we exploit the GAN architecture to find counterfactual examples for a given test sample $\mathbf{x}^{\mathbf{a},\oplus}$ and an audited model $T(\mathbf{x})$, with classes $\{\texttt{label}_1,\ldots,\texttt{label}_L\}$. Specifically, we model the counterfactual generation process as a perturbation inserted into the attribute vector $\mathbf{a}$ of the test sample, i.e. $\mathbf{b}=\mathbf{a}+\delta$, with $\delta\in\mathbb{R}^N$. This perturbed attribute vector, through G_{enc} and G_{dec}, yields a plausible image $\mathbf{x}^{\mathbf{b},\prime}$ that, when fed to the target model $T(\cdot)$, changes its predicted output. The conflict between adversarial power, plausibility and intensity of the perturbation from which the counterfactual example is produced gives rise to the multi-objective problem formulated as:

$$\min_{\boldsymbol{\delta}\in\mathbb{R}^N} f_{gan}(\mathbf{x}^{\mathbf{a},\oplus},\boldsymbol{\delta};G,D), f_{adv}(\mathbf{x}^{\mathbf{a},\oplus},\boldsymbol{\delta};G,T), f_{att}(\boldsymbol{\delta}), \tag{11}$$

where:

- $f_{gan}(\mathbf{x}^{\mathbf{a},\oplus},\boldsymbol{\delta};G,D)$ quantifies the *unlikeliness* (no plausibility) of the generated counterfactual instance through $G(\cdot)$, which is given by the difference between the output of the discriminator $D(\cdot)$ corresponding to $\mathbf{x}^{\mathbf{a},\oplus}$ and $\mathbf{x}^{\mathbf{b},\prime}$ (Wasserstein distance). The more negative this difference is, the more confident the discriminator is about the plausibility of the generated counterfactual $\mathbf{x}^{\mathbf{b},\prime}$;
- $f_{adv}(\mathbf{x}^{\mathbf{a},\oplus},\boldsymbol{\delta};G,T)$ informs about the probability that the generated counterfactual does not confuse the target model $T(\cdot)$, which is given by the negative value of the cross-entropy of the soft-max output of the target model when queried with counterfactual $\mathbf{x}^{\mathbf{b},\prime}$; and
- $f_{att}(\boldsymbol{\delta})$ measures the intensity of adversarial changes made to the input image $\mathbf{x}^{\mathbf{a},\oplus}$, which is given by $||\boldsymbol{\delta}||_2$. As we will later discuss, this measure can be replaced by other measures of similarity that do not operate over the perturbed attribute vector, but rather over the produced counterfactual image (for instance, structural similarity index measure SSIM between $\mathbf{x}^{\mathbf{a},\oplus}$ and $\mathbf{x}^{\mathbf{b},\prime}$).

To efficiently find a set of input parameter perturbations $\{\boldsymbol{\delta}\}$ balancing among the above three objectives in a Pareto-optimal fashion, we resort to multi-objective optimization algorithms. Specifically, we opt for derivative-free meta-heuristic solvers, which allow efficiently traversing the search space $\mathbb{R}^N$ of decision variables $\boldsymbol{\delta}$ and retaining progressively better non-dominated counterfactual instances without requiring any information of the derivatives of the objectives under consideration. Non-dominated counterfactual instances are a set of optimal solutions in a multiobjective counterfactual generation that strike a balance between conflicting objectives.

The multi-objective solver makes use of the already trained GAN using the weighted loss functions defined in Expressions (1) and (5) to evaluate $f_{gan}(\cdot)$, $f_{adv}(\cdot)$ and $f_{att}(\cdot)$ for a given anchor instance $\mathbf{x}^{\mathbf{a},\oplus}$ and candidate perturbation $\boldsymbol{\delta}$, so that such objective values are used to select, evolve and filter out perturbations that yield counterfactuals dominated in the Pareto space spanned by these three objectives. In other words, the multi-objective solver allows efficient traversing of the space of possible perturbations in search of counterfactuals that dominate the interplay between the aforementioned objectives. Such objectives can be evaluated for every given perturbation $\boldsymbol{\delta}$ based on the outputs of the target model $T(\cdot)$ and the $G_{enc}(\cdot)$, $G_{dec}(\cdot)$ and $D(\cdot)$ modules of the already trained GAN architecture.

Algorithm 1: Generation of multi-criteria counterfactuals.

```
Input: Target model to be audited $T(\mathbf{x})$; GAN architecture $(G, D)$; attribute classifier $C(\mathbf{x})$; annotated training set $\mathcal{D}_{train}$; test image $\mathbf{x}^{\mathbf{a},\oplus}$ for counterfactual
       study; weights $\{\lambda_i\}_{i=1}^3$
Output: Multi-criteria counterfactuals balancing between $f_{gan}(\cdot)$, $f_{adv}(\cdot)$ and $f_{att}(\cdot)$
1  Train GAN architecture via back-propagation over training dataset and loss functions in Expressions (1) to (5)
2  Initialize a population of perturbation vectors $\boldsymbol{\delta}\in\mathbb{R}^N$
3  while stopping criterion not met do
4      Apply search operators to yield offspring perturbation vectors
5      Evaluate $f_{gan}(\cdot)$ (plausibility), $f_{adv}(\cdot)$ (adversarial success) and $f_{att}(\cdot)$ (change intensity) of offspring perturbations
6      Rank perturbations in terms of Pareto optimality
7      Retain the Pareto-best perturbations in the population
8  end
9  Select non-dominated perturbations from population
10 Produce counterfactual images by querying the GAN with $\mathbf{x}^{\mathbf{a},\oplus}$ and each selected perturbation vector
```

Algorithm 1 summarizes the process of generating counterfactuals for target model $T(\cdot)$, comprising both the training phase of the GAN architecture and the meta-heuristic search for counterfactuals subject to the three conflicting objectives. The overall framework departs from the training process of a GAN architecture (line **1**) over a training dataset $\mathcal{D}_{train}$ that collects samples (images) annotated with their attribute vectors $\mathbf{a}$ (only for conditional GANs). Once trained, and similarly to the usual workflow of population-based heuristic solvers, the algorithm initializes uniformly at random a population of perturbation vectors (line **2**), which are iteratively evolved and refined (lines **3** to **8**) as per the Pareto optimality of the counterfactual images each of them produces. To this end, evolutionary search operators (crossover and mutation) are applied over the population (line **4**) to produce offspring perturbation vectors, which are then evaluated (line **5**) and ranked depending on their Pareto dominance (line **6**). By keeping in the population those perturbation vectors that score best in terms of Pareto optimality (line **7**) and iterating until a stopping criterion is met, the framework ends up with a population of Pareto-superior perturbation vectors (line **9**), that can be inspected visually by a human to understand which image components affect the most along the direction of each objective (line **10**). Indeed, since decision variables evolved by the multi-objective solvers can be directly linked to attributes imprinted to the anchor image, the amplitude of any given component of an evolved vector can be interpreted as the intensity by which the corresponding attribute is modified in the generated counterfactual. Consequently, it is possible to determine which attributes are more relevant depending on the region of the objective space under focus.

3.3. Target audiences and examples of use cases

To round up the presentation of the proposed framework, we pause briefly at the target audiences envisioned for its use, and scenarios that could illustrate its use in practical settings. Many examples could be used to exemplify these scopes, among which three specific areas currently under active investigation are chosen: bio-metric authentication, the discovery of new materials and creative industrial applications. These three use cases target two different audiences: developers and final users.

The use of biometric authentication is extensive nowadays in a manifold of sectors managing critical assets. However, auditing machine learning models used for bio-metric authentication is not straightforward. They can be audited by adversarial attack testing, but this analysis focuses on subtle (namely, not noticeable) adversarial perturbations made to an input to the audited model. Therefore, they aim at analyzing the robustness of the model against malicious attacks designed not to be easily detectable (e.g. one-pixel attacks), rather than at discerning which plausible inputs can lead to a failure of the authentication system even if not deliberately designed for this purpose. The framework presented in this work can be of great value for developers to explore the reality-bound limitations of their methods. This can help them determine complementary information requested during the process to increase the robustness of the model against plausible authentication breakpoints, which can be uncovered by examining biometric features that are especially sensitive to the adversarial power objective.

New material discovery is also a field in which high-dimensional datasets are utilized. The addition of our proposed framework might help experts reduce the amount of non-plausible composites to be synthesized, or discover diverse alternative materials with differing properties in terms of elasticity, conductivity or thermal expansion, to cite a few. This would in turn ease the practice of material experts by considerably reducing the space of possible materials to be explored and opening new possibilities in their laboratory processes without requiring any technical knowledge in Artificial Intelligence.

Finally, we highlight the possibilities brought by the proposed framework for the creative industry. Such a framework could be coupled with design software so that it would help in the generation of creative content by proposing new alternatives to already produced products (e.g. new designs of mechanical components, new audiovisual pieces, novel architectural proposals) with varying levels of compliance with respect to plausibility, amount of the change and properties that are specific to the use case at hand. In essence, the framework could be of great use for aiding the creative process hand in hand with the expert.

3.4. Limitations

Before delving into the experiments and results, it should be noted that an underlying assumption made in this study is the existence of high-quality data with concept-/attribute-based annotations that allow mapping counterfactual changes to the weighted presence of concepts in the produced counterfactual explanations. This assumption may not always hold in practice, leading to counterfactual samples with reduced interpretability. Our framework departs from the claim that explanations issued should deal with concepts and attributes, rather than with image artifacts that eventually succeed in changing the output of the audited model. The *audience* of the audited model (particularly when lacking domain-specific knowledge) requires changes to be described in terms of human-based concepts.

If high-quality data with concept-/attribute-based annotations are not available, it is not possible to properly model the distribution of plausible concept-based changes. This does not imply that counterfactual explanations cannot be produced, but the translation from input changes to induced concepts/attributes becomes much more entangled and subject to interpretation. As a matter of fact, understanding how concepts and attributes behave in the space of embedding activations in neural networks constitutes an area of very active research, yielding advances in the representation of concepts in the latent space of neural network models that can contribute to a better alignment of explanations with human-readable concepts [15,1,24].

Fortunately, the number and diversity of datasets whose samples are annotated with human-readable concepts and attributes is ever-growing in almost all domains, far more than the ones discussed in Subsection 3.3. This enables ample opportunities for showcasing the proposed framework in more use cases than the ones used in the experiments reported in this article. Furthermore, modern generative methods (e.g. pre-trained large language and stable diffusion models) can be also explored as a replacement for the GAN architectures used at the core of the framework. We will further elaborate on this in the closing section of the manuscript.

4. Experimental setup

This section introduces the actual architectures and models that were used to provide an informed response to three main research questions already mentioned in the introduction:

Q1. Is counterfactual generation an optimization problem driven by several objectives?
Q2. Do the properties of the generated counterfactual examples conform to general logic for the tasks and datasets at hand?
Q3. Do multi-criteria counterfactual explanations serve broader purposes than model explainability?

In what follows six GAN architectures are presented (Subsection 4.1), followed by six third-party classifiers (Subsection 4.2) that were audited in the six experiments designed to answer the above three questions. All GAN architectures are extracted from the literature as pre-trained cases from the original authors themselves. The classifiers are trained with the test sets of each of the GAN datasets to ensure the same data domain is maintained and no knowledge leakage is produced. Finally, details about the utilized

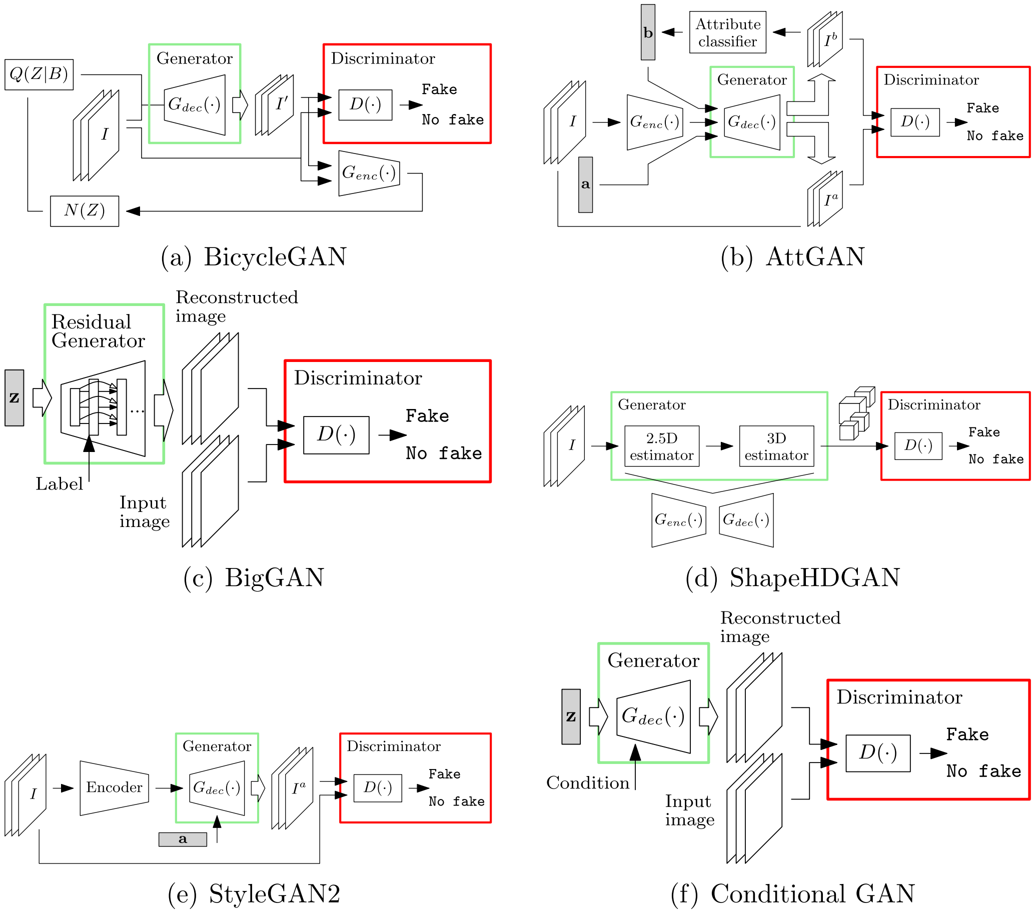


Fig. 3. Block diagram of the proposed system comprising the (a) BicycleGAN; (b) AttGAN; (c) BigGAN; (d) ShapeHDGAN; (e) StyleGAN2; and (f) Conditional GAN (CGAN).

multi-objective optimization algorithm are given (Subsection 4.3). The source code for reproducing these experiments is released at https://github.com/alejandrobarredo/COUNTGEN-Framework.

4.1. *Considered GAN architectures*

The architectures utilized fall under three main GAN categories. Although each of them consists of a particular implementation containing its particular caveats. The different GAN approaches are Conditional GAN, Unconditional GAN and a combination of both.

4.1.1. *BicycleGAN*

This first BicycleGAN architecture combines conditional and unconditional GAN architectures for the task of image-to-image translation [50]. To this end, BicycleGAN generates the output as a distribution of solutions in a conditional generative setting. The mapping is disambiguated through a latent vector which can be sampled at test time. The authors present their solution as an improvement for the known *mode collapse* problem since it reduces the pitfall of having one-to-many solutions as a result of utilizing a low-dimensional latent vector.

As shown in the diagram of Fig. 3(a), BicycleGAN combines conditional and unconditional GAN architectures to generate their own. The first part, highlighted in green, is that of a cVAE-GAN [30]: the model first encodes the ground truth into a latent space, and then it is reconstructed by means of a generator trained with a Kullback–Leibler divergence loss. The second combined model is a cLR-GAN [16]: contrarily to the first part, the cLR-GAN departs from a randomly generated latent vector, while the encoder is trained from recovering it from the output image created in the generator. Finally, the combination of these two different constraints forms the BicycleGAN architecture, which enforces the connection between the output and latent code simultaneously for both directions. This resulting architecture is able to generate more diverse and appealing images for every image-to-image translation problem.

The implementation of the network was retrieved from https://github.com/junyanz/BicycleGAN (last access: 02.11.2023) with the pre-trained models utilized for the experiments covered in the next section. The architecture consists of a U-Net generator $G(\cdot)$, which in turn contains an encoder-decoder architecture with symmetric skip connections. The discriminator $D(\cdot)$ is composed as a combination of two PatchGAN [26] of different scales which resolve the fake/real prediction for 70×70 and 140×140 image patches.

Finally, for the standalone encoder $G_{enc}(\cdot)$, a ResNet is utilized. Further information about the structure and training process of the BicycleGAN architecture can be found in [50].

4.1.2. AttGAN

This second architecture presents a conditional GAN capable of editing facial attributes of human faces while preserving the overall detail of the image [21]. In the seminal work presenting this architecture, the training process is performed by conditioning the latent vector to match the vector representing the given facial attributes for the image at hand. The network is devised such that this vector is real-valued, which allows for the inference of facial attributes for a given intensity. During inference, attributes can be changed by modifying the values of the variables in the latent vector.

Fig. 3(b) depicts a diagram of the AttGAN model, which is trained by means of two constraining conditions. For one, the model attempts to match the input attributes with the predicted attributes at the end of the architecture. For the other, the model is constrained to match the generated image to that at its input. The latter is governed by a reconstruction loss. The former, forcing the latent vector to match the attributes of the images, is governed by a standard cross-entropy loss. The combination of these two constraints results in a model capable of generating faces with varying attributes and remarkable realism.

The implementation of the network is available at https://github.com/LynnHo/AttGAN-Tensorflow (last access: 02.11.2023). The discriminator $D(\cdot)$ is composed of a stack of convolutional layers followed by fully-connected layers. The classifier $C(\cdot)$ shares all the convolutional layers from $D(\cdot)$, and follows the same structure ended in fully-connected layers. The encoder $G_{enc}(\cdot)$ is composed of several convolutional layers, while the decoder $G_{dec}(\cdot)$ is composed of a stack of transposed convolutional layers. As in BicycleGAN, a symmetrical skip connection is set between the encoder and decoder. Further information about the architecture and training process can be accessed in [21].

4.1.3. BigGAN

Despite increasing achievements in GAN modelling, the scale and diversity of datasets such as `ImageNet` have remained a hard task over the years. In [8], the authors train a generative network at the largest scale yet possible. In such research, they study the instabilities specific to the scale. They discovered the so-called *truncation trick* as a result of using orthogonal regularization to control the trade-off between the sample fidelity and variety by reducing the variance of the generator's input. This improvement allowed for a new state-of-the-art class conditional image synthesis. When trained on ImageNet at a resolution of 128×128 pixels, the newly proposed BigGAN architecture depicted in Fig. 3(c) achieves an inception score almost three times as large as that of previous image synthesis models. We use the pre-trained BigGAN implementation from https://github.com/ajbrock/BigGAN-PyTorch (last access: 02.11.2023).

4.1.4. ShapeHDGAN

This fourth ShapeHDGAN architecture is capable of rendering 3D meshes of objects from single 2D views. This particular task is of great complexity given that the solution landscape is composed of countless shapes that do not pertain to an object and renders them implausible. Most existing approaches fail to generate detailed objects. ShapeHDGAN gives a solution to this problem by virtue of a generative environment with adversarially learned shape priors that serve the purpose of penalizing if the model renders unrealistic meshes.

As shown in Fig. 3(d) the model consists of two main components. A 2.5D sketch estimator and a 3D shape estimator that predicts a 3D object from an image. It consists of three stages. In the first stage, the 2.5D estimator – an encoder-decoder structure – predicts the object depth, normals and silhouette from an RGB image. Then, the second stage generates a 3D shape from the previous 2.5D sketch. The last stage is composed of an adversarially trained CNN that tunes the generated shape into a real object.

The implementation was retrieved from https://github.com/xiumingzhang/GenRe-ShapeHD (last access: 02.11.2023). The 2.5D sketch estimator is composed of a ResNet-18 encoder $G_{enc}(\cdot)$ mapping a 256×256 image into 512 feature maps of size 8×8. The $G_{dec}(\cdot)$ model has four stacked transposed convolutional layers. The predicted silhouette permits to mask off the depth and normal estimations to be then used as the input of the 3D generator. The 3D shape estimator is also composed of an encoder-decoder architecture. The encoder is adapted from a ResNet-18 to handle 4 channels and encode them into a 200-dimensional latent vector. The decoder comprises five stacked transposed convolutional layers, which generate a $128 \times 128 \times 128$ voxel at its output. Further details are available at [48].

4.1.5. StyleGAN2

StyleGAN [28] is an unconditional GAN architecture with one of the most realistic results for unconditional generative image modelling. For this study, we choose the StyleGAN2 implementation, which is a revised variant of the original StyleGAN model with small albeit intelligently devised modifications to the generator model. The implementation was retrieved from https://github.com/NVlabs/stylegan2 (last access: 02.11.2023) with the pre-trained models for the experiments carried out in the following sections.

4.1.6. Conditional GAN

Finally, we decided to add a last model that allows us to explore some variations within. This time, we selected a well-known conditional GAN architecture [34] trained over the MNIST image classification dataset. The conditional GAN departs from a random noise vector and a single variable that acts as a condition for the generation process. In this way, the generative network learns to switch between the learned distributions for each label by means of an input condition. This feature resembles that of AttGAN, with the difference that in this one, the models do not start from an encoding.

Table 2
Structure and training parameters of the models audited by the proposed framework.

GAN	Audited classifier $T(\mathbf{x})$	
	Network architecture	Training parameters

BicycleGAN	Conv2d(64, 3 × 3, ReLu) + Conv2d(32, 3 × 3, ReLu)
	+ Dense(1, Sigmoid)
		Adam, binary cross-entropy loss

AttGAN	Conv2d(16, 3 × 3, ReLu) + Dropout(0.1)
	+ Conv2d(4, 3 × 3, ReLu) + Dense(1,Sigmoid)
		Adam, binary cross-entropy loss

BigGAN	Conv2d(64, 7 × 7, ReLu) + MaxPooling(3 × 3)
	+ Conv2d(64, 3 × 3, ReLu) + Conv2d(128, 3 × 3, ReLu)
	+ Conv2d(256, 3 × 3, ReLu) + Conv2d(512, 3 × 3, ReLu)
	+ AvgPooling(7 × 7) + Dense(500,1000) + Dense(1000,Softmax)
		SGD(0.01, 0.9), categorical cross-entropy loss

ShapeHDGAN	Conv2d(32, 3 × 3, ReLu) + BatchNorm + MaxPooling(2 × 2) + Conv2d(8, 3 × 3, ReLu) + BatchNorm + MaxPooling(2 × 2) + Dense(100,ReLu) + Dense(1, Sigmoid)	SGD(0.01, 0.9), binary cross-entropy loss

StyleGAN2	Conv2d(16, 3 × 3, ReLu) + Dropout(0.1)
	+ Conv2d(4, 3 × 3, ReLu) + Dense(1,Sigmoid)
		Adam, binary cross-entropy loss

cGAN	Conv2d(32, 3 × 3, ReLu) + BatchNorm
	+ MaxPooling(2 × 2) + Dense(100,ReLu)
	+ Dense(10,SoftMax)
		SGD(0.01, 0.9), categorical cross-entropy loss

Conv2d(A,B,C): convolutional layer with A filters of size B and activation C.
SGD(l,m): Stochastic Gradient Descent with learning rate l and momentum m.
In all cases the batch size is set to 16 instances, and the number of epochs is 10.
Flattening operations are not displayed for clarity.

Fig. 3(f) shows the structure of the conditional GAN model. The implementation was retrieved from the public python library GANS2 (https://github.com/tlatkowski/gans-2.0, last access: 02.11.2023), which includes a set of ready-to-build, plug-and-play GAN architectures.

4.2. Audited classification models

After introducing the GAN models under consideration, we now introduce the models that will be audited by means of our GAN-based counterfactual generation framework. For the experiment utilizing BicycleGAN, a classifier is trained to predict the type of footwear corresponding to the image fed at its input (`Shoe` versus `NoShoe`). For the case considering AttGAN, the classifier to be audited predicts whether the human face input to the model corresponds to a `male` or to a `female`. The case using BigGAN considers a ResNet-50 classifier that discriminates among the 1000 classes represented in the `ImageNet` dataset. When the framework considers ShapeHDGAN, the classifier is trained to distinguish between a `chair` and a `Xbox`. For StyleGAN2, the classifier discriminates whether the input image is a `cathedral` or an `office`. Finally, the classifier audited by our framework configured with the cGAN aims to address a multi-class classification problem over the same MNIST dataset, yet ensuring that different data partitions are used for training the cGAN model and the classifier itself.

These third-party models consist of several convolutional and residual layers, ending in a series of fully connected layers that connect the visual features extracted by the former with the categories defined in the dataset under consideration. Every classifier model was trained with the test data that was not used for training the corresponding GAN architecture, thereby ensuring no information leakage between the generators and the third-party models to be audited. Table 2 summarizes the topological configuration of the models for which counterfactuals are generated by our framework, as well as the training parameters set for every case.

The performance achieved by the trained classifiers over a 20% holdout of their dataset is reported in Table 3, together with the number of classes, total examples to train and validate the audited model, and the class balance ratio. As can be observed in this table, the audited models reach a very high accuracy (over 94% in most cases, except for the `ImageNet` classifier due to the notably larger number of classes of the dataset), so that the adversarial success of the produced counterfactual examples can be rather attributed to the explanatory capabilities of the devised framework than to a bad performance of the audited classifier. Furthermore, we verified the capability of the trained GANs to conditionally generate new instances based on a vector of attributes, by 1) verifying the convergence of the discriminator and classifier loss function over the training epochs, and 2) visually inspecting the quality of several test instances and arbitrary perturbations.

4.3. Multi-objective optimization algorithm

We recall that the optimizer is in charge of tuning the output of the GAN generator to 1) maximize the difference in the result of the audited classifier (adversarial power); 2) minimize the number of changes induced in the produced counterfactual parameters (change intensity), and 3) maximize the Wasserstein distance between the real and fake examples (plausibility).

The search for counterfactual instances optimally balancing among these objectives can be efficiently performed by using a multi-objective evolutionary algorithm. Among the multitude of approaches falling within this family of meta-heuristic solvers, we select

Table 3
Dataset and accuracy of the different black-box classifiers under target.

GAN	Dataset	# Examples	Classes	Class Balance	Accuracy	Source
BicycleGAN	`Edges2Shoes`	300	2	45%/55%	94%	[26]
AttGAN	`CelebA`	900	2	49%/51%	98%	[32]
BigGAN	`ImageNet`	1,281,167 (training)	1000	Varying	75%	[20]
ShapeHDGAN	`ShapeNet`	600	2	49%/51%	96%	[10]
StyleGAN2	`Style`	540	2	49%/51%	98%	[49]
cGAN	`MNIST`	9000	10	10% each	96%	–

NSGA2 with a population size of 100 individuals, 100 offspring produced at every generation, polynomial mutation with probability $1/N$ (with N denoting the number of decision variables, which vary depending on the experiment and GAN under consideration) and distribution index equal to 20, SBX crossover with probability 0.9 and distribution index 20, and 50 generations (equivalent to 5000 evaluated individuals per run). The use of this optimizer allows for a genetic search guided by non-dominated sorting in the selection phase, yielding a Pareto-dominant set of counterfactual examples that constitute the output of the framework. For its implementation we rely on the jMetalPy library for multi-objective optimization [6].

5. Results and discussion

Answers to each of the research questions posed previously will be summarized after the analysis of the produced counterfactual explanations for each of the six audited models detailed in Table 2. Results of these six experiments are given in Subsections 5.1 to 5.6, which should be identically interpreted as described in the next two paragraphs.

Presentation and interpretation of the results For every experiment, we draw at random one anchor image $\mathbf{x}^{\mathbf{a},\oplus}$ from the test partition of the audited model and inspect the produced set of counterfactual variants both visually and quantitatively as per the three objectives under consideration. This examination of the results will be arranged similarly across experiments, portraying the output of the framework in a three-dimensional plot comprising the Pareto front approximated by the multi-objective solver. Each of the axes of this plot is driven by one of such objectives: change intensity $f_{att}(\cdot)$, adversarial power $f_{adv}(\cdot)$ and plausibility $f_{gan}(\cdot)$, all defined in Subsection 3.2. It is important to note that for easing the visualization of the fronts, plausibility and adversarial power are inverted by displaying $1-f_{gan}(\cdot)$ and $1-f_{adv}(\cdot)$, so that $1-f_{gan}(\cdot)\geq 0.5$ denotes the region over which the counterfactual can be considered to be plausible. Similarly, the higher $1-f_{adv}(\cdot)$ is, the larger the difference between the outputs of the audited model when fed with the anchor image $\mathbf{x}^{\mathbf{a},\oplus}$ and its counterfactual variant will be (larger *adversarial power*).

Visualization of the counterfactual explanations In the depicted Pareto front approximations for every experiment (in the form of three-dimensional scatter plots, parallel lines visualizations and chord diagrams), several specific counterfactual examples scattered over the front are highlighted with coloured markers. These markers refer to the images plotted on the last subplot of each figure so that it is possible to assess the counterfactual image/voxel corresponding to each of such points. The first image shown in the top row of images shown on the right of the figure represents the reference (anchor) image $\mathbf{x}^{\mathbf{a},\oplus}$, which is the departing point for the counterfactual generation. The first image shown in the bottom row of images is always the image belonging to the opposite class (or a targeted class in the case of the MNIST dataset) whose soft-max output corresponding to its class is the lowest (worst predicted example of the other class existing in the dataset). Below every image, a bar diagram can be observed representing the value of the objectives corresponding to the image at hand.

We now proceed with a detailed discussion of every experiment:

5.1. Experiment #1: BicycleGAN-based counterfactual generation for auditing a `Shoe` versus `NoShoe` footwear classifier

The outcomes of this first experiment are shown in Fig. 4.a to Fig. 4.d. The first one depicts a three-dimensional scatter plot of the Pareto front approximation generated by our framework for the man shoe selected for the anchor image $\mathbf{x}^{\mathbf{a},\oplus}$. Figs. 4.b and 4.c correspondingly depict the parallel lines plot and chord plot of the counterfactual examples, which not only ease the visual inspection of the objective ranges spanned by the Pareto front but also show the diversity sought for the counterfactual examples selected from the front. A colour correspondence is fixed across subplots for the reader to track each counterexample through them. In Fig. 4.d a colouring pattern is distinguishable over all the counterfactual examples highlighted in the approximated front. The image of the man shoe that serves as the anchor image $\mathbf{x}^{\mathbf{a},\oplus}$ appears to be complete. However, the original colours are removed and uniformized all over the image. This fact informs about the influence of the colour on the predicted label of the model.

Returning to our intuition exposed in Section 3, two concerns must be kept in mind when analyzing these results. First, understanding the constraints of the dataset in use is of paramount importance. The data with which the classifier was trained is composed of different footwear instances. However, this dataset accounts just for a limited subset of the different possible footwear instances available in reality for both classes. This fact will make the predictions of the classifier change sharply between one class and the other when the instance for which it is asked does not conform to the class-dependent distribution of the training dataset. The second concern refers to the spread in the prediction scores. The solution front is depicted in Figs. 4.a to 4.c shows a nice spread in the

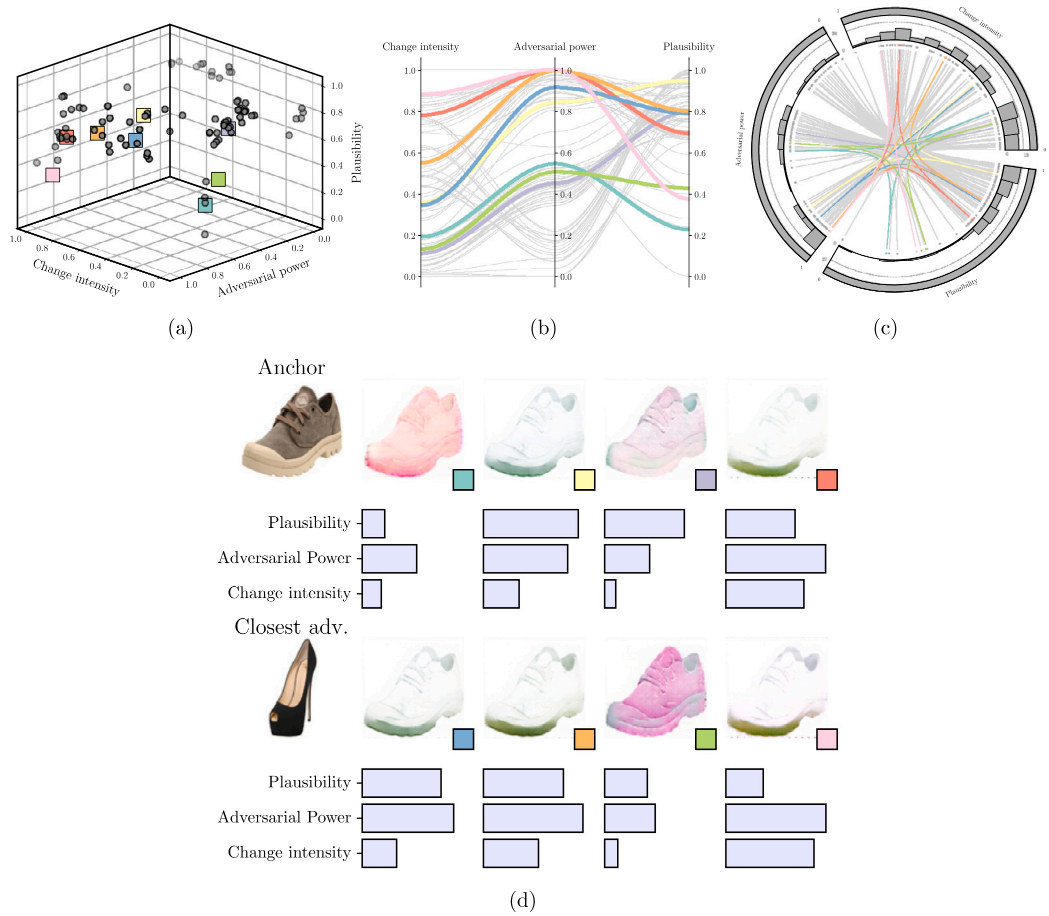


Fig. 4. (a) Pareto front of the counterfactual examples generated for a `Shoe` example by the proposed framework configured with a BicycleGAN model; (b) parallel lines visualization of the Pareto front; (c) chord plot of the Pareto front; (d) produced counterfactuals, together with the anchor image and the closest image of the opposite class for which the counterfactual is made. We note the colour correspondence across all these subplots, by which the reader can follow the position of any of the highlighted counterfactual instances in the front, its objective values and image information.

prediction scores at first glance. However, this spread of solutions in the objective space does not entail that the corresponding counterfactual instances are visually diverse. We depict just 8 out of the 100 solutions in the approximated Pareto front, but they suffice to showcase that every generated counterfactual is very similar to each other with the exception of colour. This suggests that the classifier is very susceptible to the colour feature and that the shape of the footwear is so relevant to the task that the counterfactual generation process needs to retain this feature to ensure plausibility. This bias is one of the insights provided by the proposed framework in this first experiment.

The aforementioned statement is supported by Fig. 5, which depicts the mean luminance of RGB pixels averaged over all the training examples of every class used for the audited model. Luminance has been computed as:

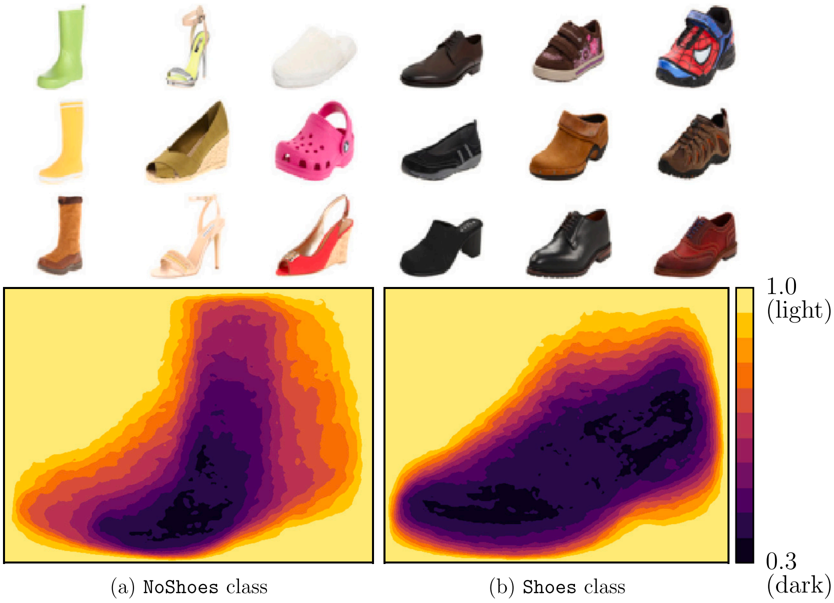
$$\ell(RGB) = (0.2126 \cdot R + 0.7152 \cdot G + 0.0722 \cdot B)/255, \tag{12}$$

where $\ell(RGB) \in \mathbb{R}[0,1]$ denotes a measure of luminance (0: dark, 1: light) of a pixel with R (red), G (green) and B (blue) channel values. As it can be observed in the bottom left plot of this figure, `shoe` instances have a clear bias in terms of footwear shape and image orientation, whereas the central part of the footwear for both classes is darker than the background. This is the reason why our proposed framework operates exclusively on the colour feature and maintains the shape of the footwear when attempting to produce a counterfactual example for a `shoe`, yielding differently (brighter) coloured yet identically shaped variants of the anchor.

5.2. Experiment #2: AttGAN-based counterfactual generation for auditing a `Man` versus `Woman` gender classifier

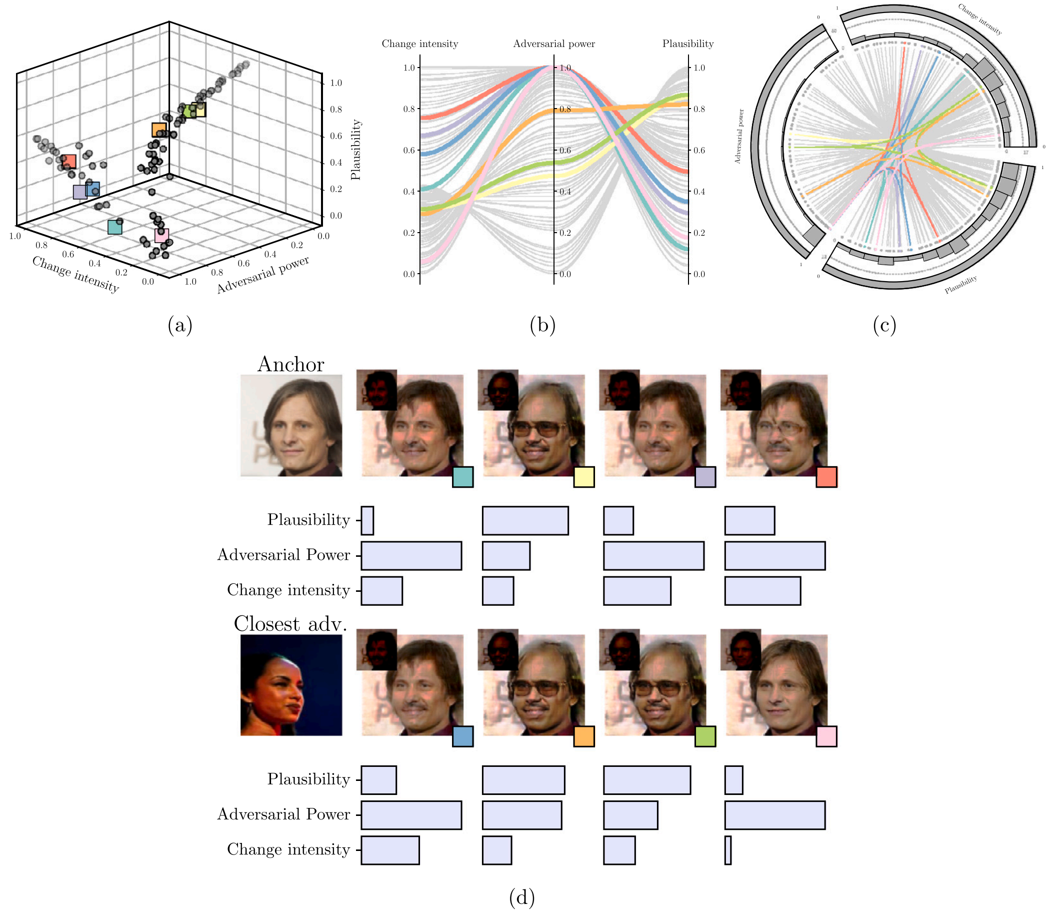
The outcome of the devised framework corresponding to this second experiment is shown in Figs. 6.a to 6.d. In this case, the reference image $\mathbf{x}^{\mathbf{a},\oplus}$ is an instance of the `Man` class from the `CelebA` dataset.

We begin by inspecting the shape of the produced Pareto front approximation in Figs. 6.a to 6.c. It can be observed that diverse counterfactual explanations are found in the trade-off between plausibility and adversarial power, as well as between the intensity of the change and plausibility. Interestingly, adversarial power and change intensity seem to be less conflicting with each other, as



(a) NoShoes class (b) Shoes class

Fig. 5. Analysis of the average RGB luminance $\ell(\boldsymbol{RGB})$ of the Shoe vs NoShoe dataset used to train the target classifier for the BicycleGAN experiment, together with some few examples of every class.



(a) (b) (c) (d)

Fig. 6. (a) Pareto front; (b) parallel lines visualization; (c) chord plot; and (d) images of the counterfactual examples generated for a male example by the proposed framework configured with an AttGAN model. Every nested couple of images in subplot (d) represents the original produced counterfactual (small image located in the upper left corner of every plot) and its contrast-adjusted version using histogram equalization.

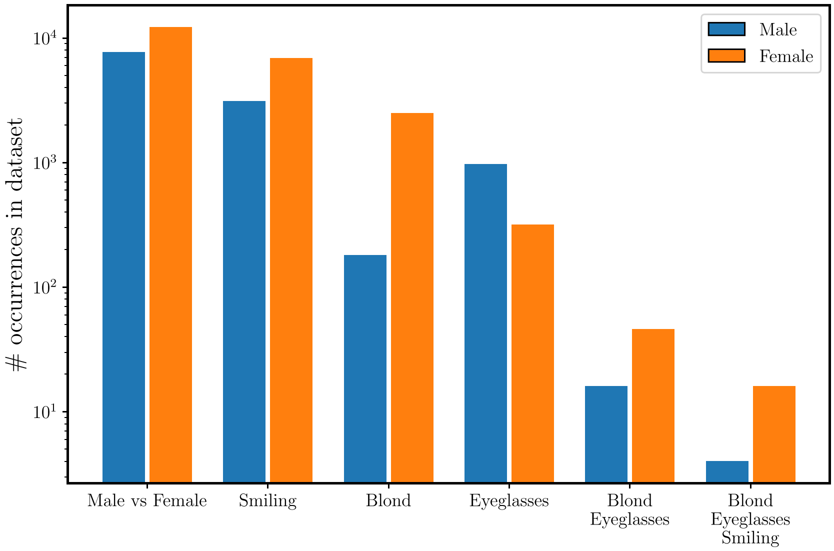


Fig. 7. Diagram showing the occurrence within the training dataset of the audited model of different feature combinations, split between `male` and `female` examples.

no counterfactuals with high change intensity (> 0.5) and low adversarial power were produced by the framework. The reason for the behaviour of these two objectives may reside in the characteristics of the dataset and GAN in use: a high perturbation in the attribute vector imprints already enough changes in the generated counterfactual image to mislead the audit classifier, at the cost of degradation of their plausibility. Similarly, counterfactuals with small change intensity can achieve a wide diversity of adversarial power values, which reveals that some perturbations are more effective than others when inducing a change in the output of the target model. However, it is clear (especially from the parallel lines visualization in Fig. 6.b) that those small changes that have high adversarial power, in general, are not plausible.

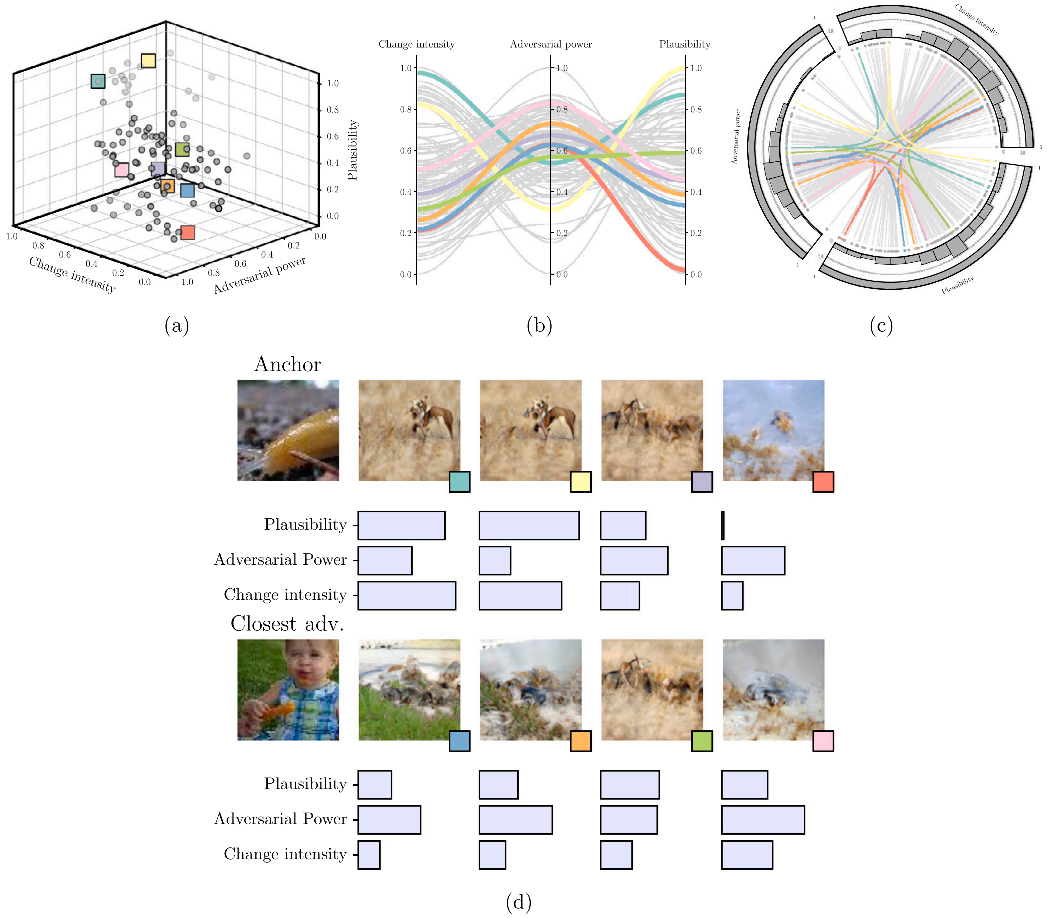
When qualitatively examining the generated counterfactuals, the plots nested in Fig. 6.c reveal that once again, the luminance is a deciding factor for adversarially modifying the anchor image. Leaving aside modifications over the colour space, it is important to note that the plausibility of counterfactuals seems to be tightly linked to the insertion of glasses or a smiling pose. On the contrary, counterfactuals that produce an intense drift towards the `Female` class in the audited classifier insert long blonde hair into the anchor image. In this experiment, these patterns are related to the constraints imposed by the dataset. However, differently from the previous experiment, the produced counterfactuals are not exiting the data domain over which the model was trained, but are rather exploiting biases existing in the data. Most counterfactuals seen in the front have blonde hair, glasses or a smile pose, whether alone or combined.

In order to explore the reason for such a recurring set of counterfactual features, Fig. 7 depicts bar diagrams showing the differences in terms of occurrence over the training examples of different combinations of the three attributes, differentiating between counts corresponding to the `male` and `female` classes. It is straightforward to note that the majority of examples featuring any of the combinations of these three attributes belong to the `female` class. Given a face, if it contains those three attributes, it is quite probably a female. This conclusion is also supported by the fact that the proportion of `male` instances wearing eyeglasses is notably higher than that of `female` examples; however, when considering eyeglasses together with the other two attributes, the proportion changes in favour of `female` examples. This is the reason why, among the counterfactual instances shown in Fig. 6.c, those using *Eyeglasses* to turn actor Viggo Mortensen into a woman imply changing his hair colour to blonde and modifying his expression to include an open smile. In summary: counterfactual instances can help unveil biases in the training data that otherwise could pass unnoticed and could affect the generalization of the target model.

5.3. Experiment #3: BigGAN-based counterfactual generation for auditing an `ImageNet` classifier

This third experiment is devised to exemplify that the proposed framework can be used to produce counterfactuals in complex tasks comprising a higher number of classes. To this end, as has been mentioned in Section 4 we resort to a ResNet18 classifier trained over `ImageNet`. For the sake of brevity, we will discuss the set of counterfactual examples generated for an anchor image belonging to class `Fiddler Crab`, using class `pyjama, pyjama, pj's, jammies` (hereafter, `Pyjama`) as the target label driving the adversarial modifications imprinted to the anchor image.

Results obtained for this third experiment are collected and shown in Figs. 8.a to 8.d. It must be first noted that one could intuitively expect that the plausible transformation of an image belonging to class `Fiddler Crab` into a counterfactual belonging to class `Pyjama` should be difficult to achieve, due to the visual differences that naturally arise between both classes. Thereby, if plausibility is kept as one of the objectives driving the evolution of counterfactuals, *colour information* should play an important role in adversarial power, whereas *shape information* would conversely act as a guarantee of plausibility. This is indeed what can be observed



(a) (b) (c)

(d)

Fig. 8. (a) Pareto front; (b) parallel lines visualization; (c) chord plot; and (d) images of the counterfactual examples generated for a `Fiddler Crab` example by the proposed framework configured with a BigGAN model.

in the results. On one hand, the Pareto front does not contain counterfactuals that achieve a large adversarial power and are clearly plausible, evincing the significant separation between the two classes. On the other hand, the visualization of the counterfactual images in Fig. 8.d aligns with the aforementioned intuition: highly plausible counterfactuals retain some image artefacts (e.g., crab legs) that are typical of the class of the anchor image, whereas colour information matches that of the closest adversarial of the target label `Pyjama`. This is also supported by the fact that in unlikely counterfactuals (e.g., 🟥, 🟦, 🟧 and ◻) the only visual aspect that ties the image content to the target label is colour information (green, blue and light brown matching the colours present in the closest adversarial), without any discernible image information that could improve the plausibility of the counterfactual.

5.4. Experiment #4: ShapeHDGAN-based counterfactual generation for auditing a `Chair` versus `Xbox` voxel classifier

This fourth case of the devised set of experiments comprises an audited classifier that discriminates whether the voxel at its input is a `chair` or a `Xbox`. Therefore, it operates over three-dimensional data, increasing the complexity of qualitatively evaluating the produced counterfactuals with respect to previous experiments.

The results elicited for a `chair` target instance $\mathbf{x}^{\mathbf{a},\oplus}$ are shown in Figs. 9.a to 9.d. A first inspection of the counterfactual voxels highlighted in the approximated Pareto front suggests that it is hard to analyze what the audited classifier observes in these inputs to get fooled and predict a `Xbox`. It appears that a dense middle part is capable of misleading the classifier. Voxels being generated by the framework resemble a chair but possess a clearly more dense middle part. It is quite revealing to see how a chair and an Xbox can be of any resemblance. Interestingly, a concern to bring up here is that of scale. These voxels (both the generated ones and the dataset over which the ShapeHDGAN model was trained) are normalized, which in turn means that the size of the object has been lost. Scaling can be an interesting option to improve the resolution of small objects. In this case, however, it can be the reason to make this classifier prone to error. See Fig. 10.

This last statement can be buttressed by analyzing which structural parts of the counterfactual voxel are of the highest importance for the audited classifier to produce its prediction. This can be done by resorting to gradient-based local post-hoc explanation methods such as Grad-CAM++ [11]. As can be seen in the examples depicted in this figure, most of the observational focus of the model is placed on the vertical rectangular part of the `chair`, which conforms to intuition given that the actual shape of a `Xbox` is rectangular.

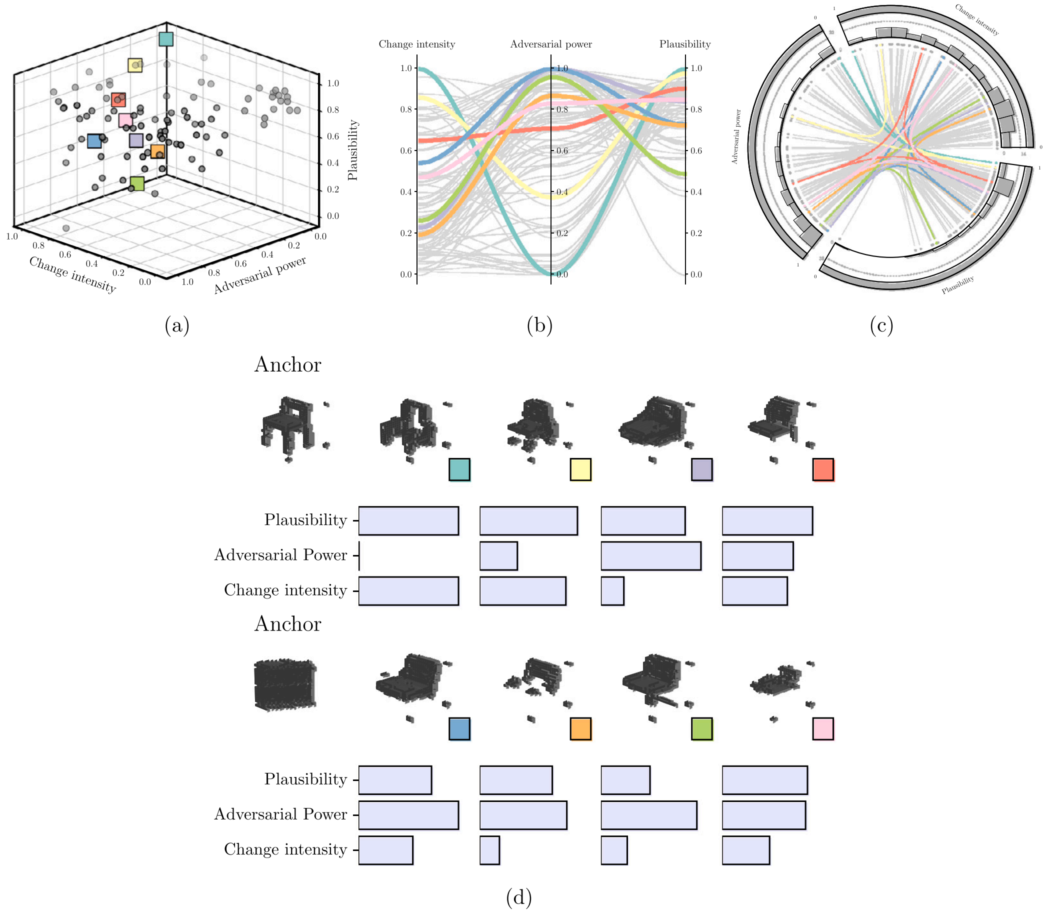


Fig. 9. (a) Pareto front; (b) parallel lines visualization; (c) chord plot; and (d) images of the counterfactual examples generated for a `chair` example by the proposed framework configured with a Shape3DGAN model.

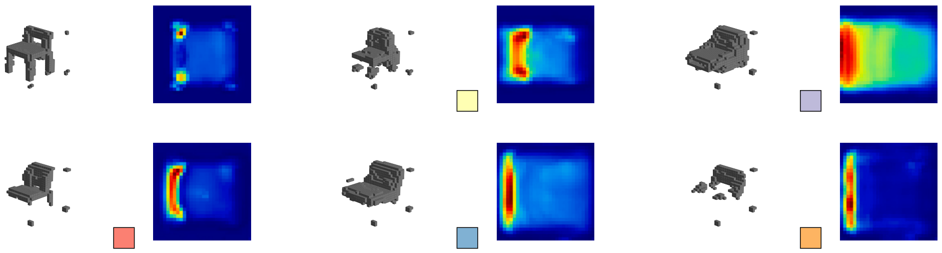


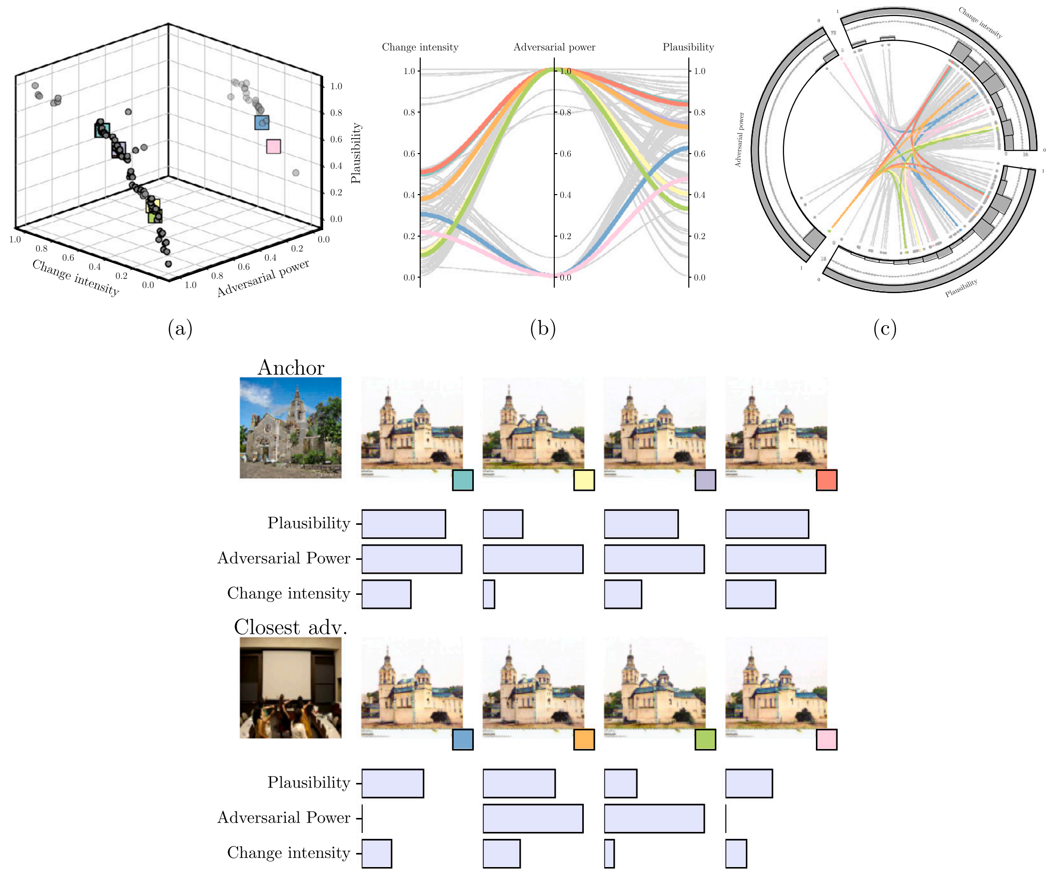
Fig. 10. Local explanations (heatmaps via Grad-CAM++) corresponding to the anchor voxel (leftmost pair of images) and two of the counterfactuals depicted in Fig. 9.

Therefore, counterfactuals for a `chair` instance wherein the vertical part (*backrest*) is reinforced can bias the audited model without jeopardizing their plausibility.

5.5. *Experiment #5: StyleGAN2-based counterfactual generation for auditing a* `Cathedral` *versus* `Office` *classifier*

The results of this experiment (Figs. 11.a to 11.d) unveil a link between the luminance of the overall counterfactual image and its ability to mislead the model. However, at this time the spread of counterfactuals over the adversarial power dimension of the Pareto front is notably lower than in the previous experiments. If the results are compared with those of Fig. 4, in this case, the missing spread observed in the front is validated with what can be visually discerned in the counterfactuals given by the framework.

Following this last observation, we focus on the analysis of the visual differences between the anchor image and the produced counterfactuals shown in Fig. 12. Specifically, the plot shows the heatmap of mean absolute differences (averaged over the RGB channels) and the SSIM (Structural Similarity Index Measure) among the original anchor image and its counterfactual version. As can be inferred from the visuals in the first row of the figure, the framework exploits a burned-out background with minor structural



(a) (b) (c)

(d)

Fig. 11. (a) Pareto front; (b) parallel lines visualization; (c) chord plot; and (d) images of the counterfactual examples generated for a `Cathedral` example by the proposed framework configured with a StyleGAN2 model.

changes in the image. This seems reasonable with the spread found in the front: it shows that these changes in the background of the image can completely fool the model, but there are no changes that would account for a well-spread front since the structural differences among both classes are large. Furthermore, `cathedral` instances undergo a misrepresentation bias in the dataset, in the sense that none of the `cathedral` training examples has a totally overcast sky. This suggests that whitening the background of the image may grant a chance for the counterfactual to mislead the audited classifier, yet without any guarantee for success given the scarce similarity between images belonging to both classes.

5.6. Experiment #6: CGAN-based counterfactual generation for auditing a `MNIST` classifier

In correspondence to Q3, this last experiment is devised to elucidate whether the output of the proposed framework can be used for any other purpose than the explainability of the target model. To this end, we run and assess visually the counterfactuals generated for the digit classification task defined over the well-known `MNIST` dataset. The characterization of every class defined in this dataset is done by a naive conditional GAN.

Figs. 13.a to 13.d portray the output of the framework when generating counterfactuals for an anchor image $\mathbf{x}^{\mathbf{a},\oplus}$ corresponding to digit 4. From what can be observed in the samples extracted from the front, visual information corresponding to digits 4 and 8 appears to be interfering with the capability of the audited model to discriminate among them. This intuition is buttressed by the fact that the closest element is a sample corresponding to digit 8, as displayed in the first bottom image of Fig. 13.d. Indeed, once again misrepresented visual artefacts in the dataset are opening a path to generate plausible counterfactuals, since most instances generated by the framework are digits with incomplete shapes. This may come from the fact that the MNIST dataset is mostly composed of digits that are correctly written.

We prove the converse to this statement by running again the experiment with an additionally inserted class in the dataset that contains digits of every class over which a part has been erased. This narrows the opportunities for the framework to generate counterfactuals by erasing selected shape fragments of the anchor digit. This is confirmed in Figs. 14.a to 14.d, which depict the output of the framework in this alternative setting: the counterfactual instances generated can be declared to be plausible with respect to this extended dataset, yet a visual inspection of their corresponding digits concludes that they do not resemble a numerical

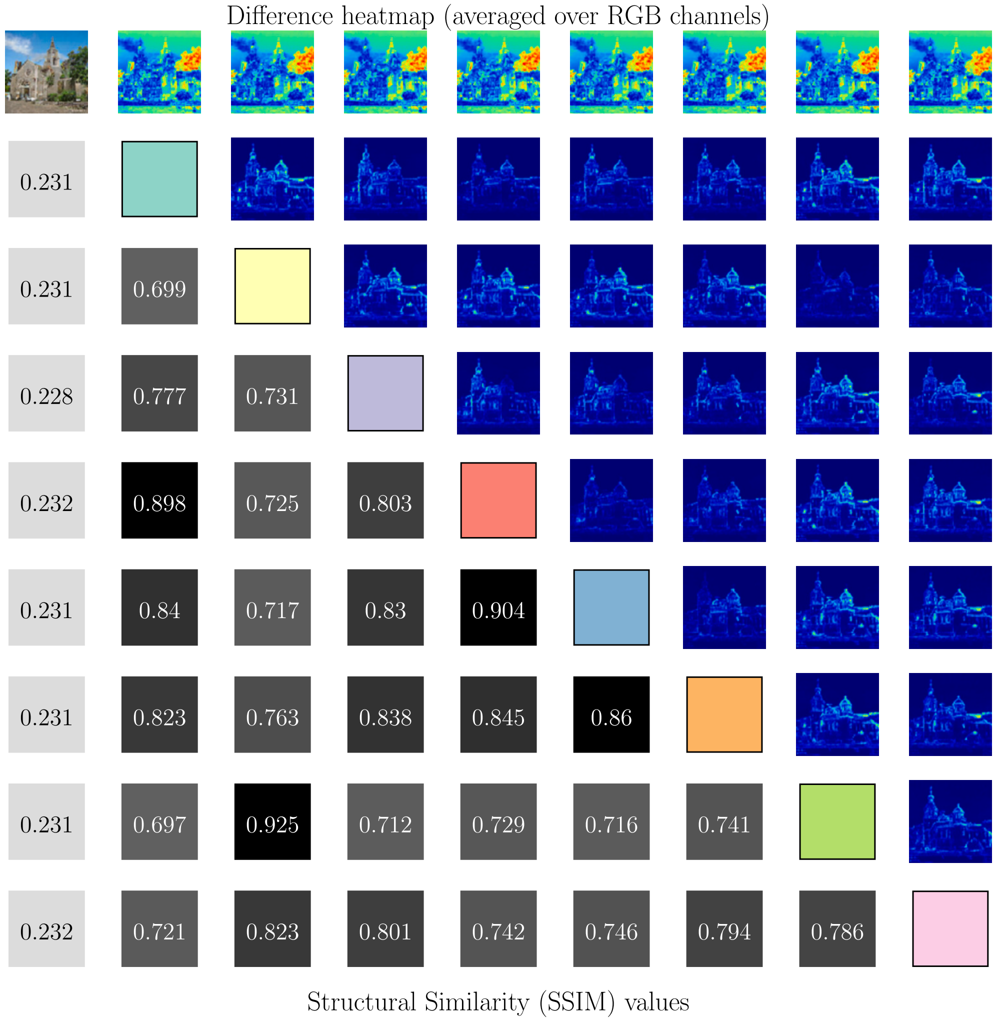


Fig. 12. Comparison between the original counterfactuals and the anchor image following the coloured markers shown previously in Fig. 6: the upper triangular part of the matrix is composed of heatmaps depicting the mean absolute difference of the RGB pixels of every pair of images in comparison, whereas the lower triangular part denotes the SSIM value quantitatively reflecting the similarity between the images.

digit. In summary: the output of our framework can tell which domain over the image (colour, shape) can be leveraged to make the audited model more robust against input artefacts.

6. Conclusion

This manuscript has proposed a novel framework that leverages the generative strength of GAN architectures and the efficient exploration capabilities of multi-objective optimization algorithms to traverse search spaces of large dimensionality. Our research hypothesis departs from the need to inform the user with further information beyond the capability of the counterfactual instances to adversarially change the output of the black-box model under study. Counterfactual generation approaches must ensure that the change *can occur*, and inform the user about the intensity of the change of the generated counterfactual, and the severity by which the output of the model would change should the counterfactual actually occur.

The devised framework builds upon this hypothesis to produce multi-criteria counterfactual explanations for a given input example and a black-box model to be audited. Specifically, a GAN model is used to furnish a generative model that characterizes the distribution of input examples which, together with its discriminator module and its conditional dependence on an attribute vector, synthesizes examples that can be considered *plausible*. The trained GAN is therefore used as a proxy evaluator of the plausibility of new data instances that change the output of the audited model (counterfactual explanations). Our designed framework seeks the set of counterfactual examples that best balance between *plausibility*, and *adversarial power*, incorporating a third objective (*change intensity*) that may be also in conflict, depending on the dataset at hand. An analysis of the limitations and assumptions made in the present study has been also done, highlighting the need for attribute-annotated data for the framework to properly model the distribution of concept-based modifications spanned by the input domain of the audited model.

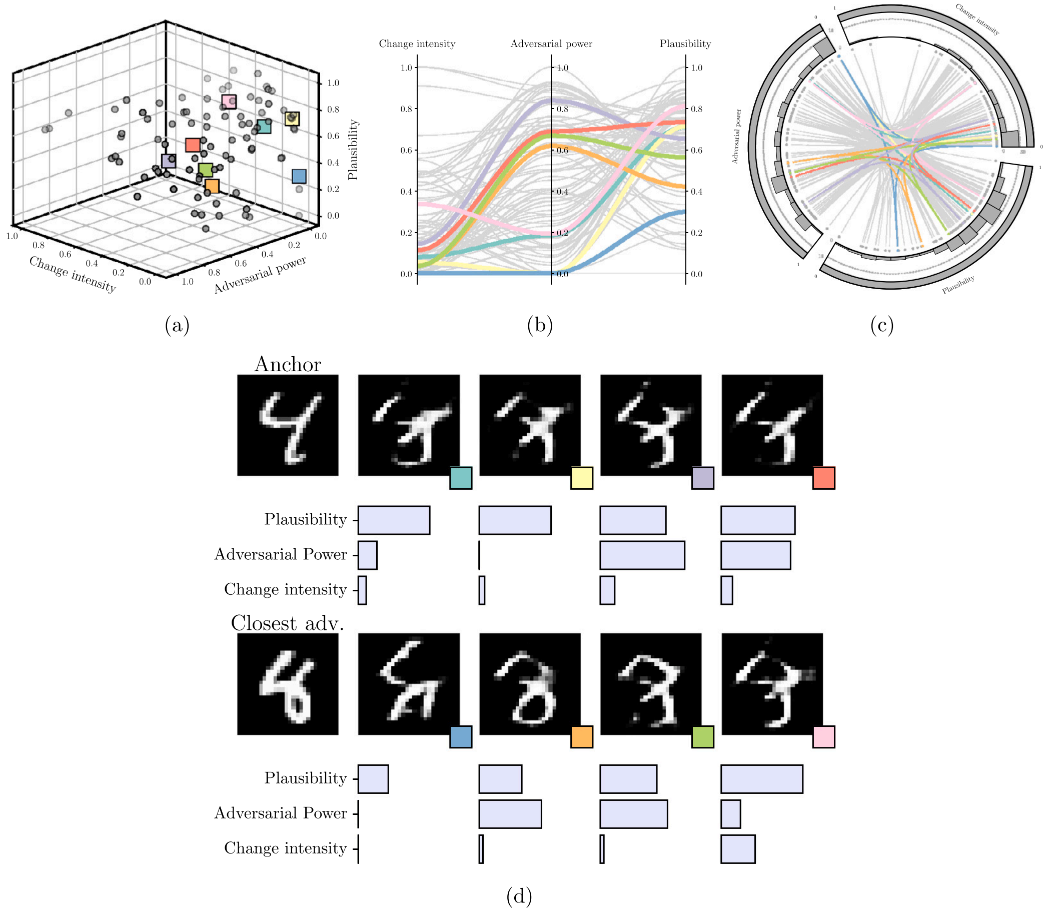


Fig. 13. (a) Pareto front; (b) parallel lines visualization; (c) chord plot; and (d) images of the counterfactual examples generated for an MNIST digit classification model by the proposed framework configured with a CGAN model.

Six experiments have been run and discussed to answer three research questions aimed at understanding the contribution of the framework to the explainability and understanding of the model being audited. The conclusions drawn with respect to such questions are given below:

- *Q1. Is counterfactual generation an optimization problem driven by several objectives?*
  As evinced by the Pareto front approximations obtained for the six experiments, counterfactual explanations are clearly governed by multiple objectives of relevant importance for the search. Depending on the dataset, some of such objectives could not be conflicting with each other. Nevertheless, the task of finding good counterfactual explanations must be approached as a search comprising different goals for the sake of a more enriched interface for the user of the audited model.
- *Q2. Do the properties of the generated counterfactual examples conform to general logic for the tasks and datasets at hand?*
  Definitely: our discussion on the results obtained for every experiment we have qualitatively inspected images and voxel volumes corresponding to the produced counterfactual instances. Artefacts observed in such adversarial images not only can be explained as departing from common sense as per the task addressed by the audited model (e.g. colour variations or emphasized structural parts of the voxels) but also exploit differences and similarities found among the data classes feeding the model at hand.
- *Q3. Do multi-criteria counterfactual explanations serve broader purposes than model explainability?*
  Indeed, the counterfactual analysis may contribute to the discovery of hidden biases resulting from misrepresentations in the training dataset of the audited model. Our discussions have empirically identified that counterfactual explanations can reflect such misrepresentations which, depending on the context, can be understood as a hidden compositional (attribute-class) bias or a potential vulnerability for adversarial attacks.

Our framework presented in this work has showcased that counterfactual explanations must be tackled as a multi-faceted challenge due to the diversity of audiences and profiles for which they are generated. Understanding how a black box behaves within the prediction boundaries of its feature spaces empowers non-expert users. Furthermore, ensuring that counterfactual changes can be mapped to the presence of concepts and attributes (as guaranteed by our framework whenever high-quality annotated data are

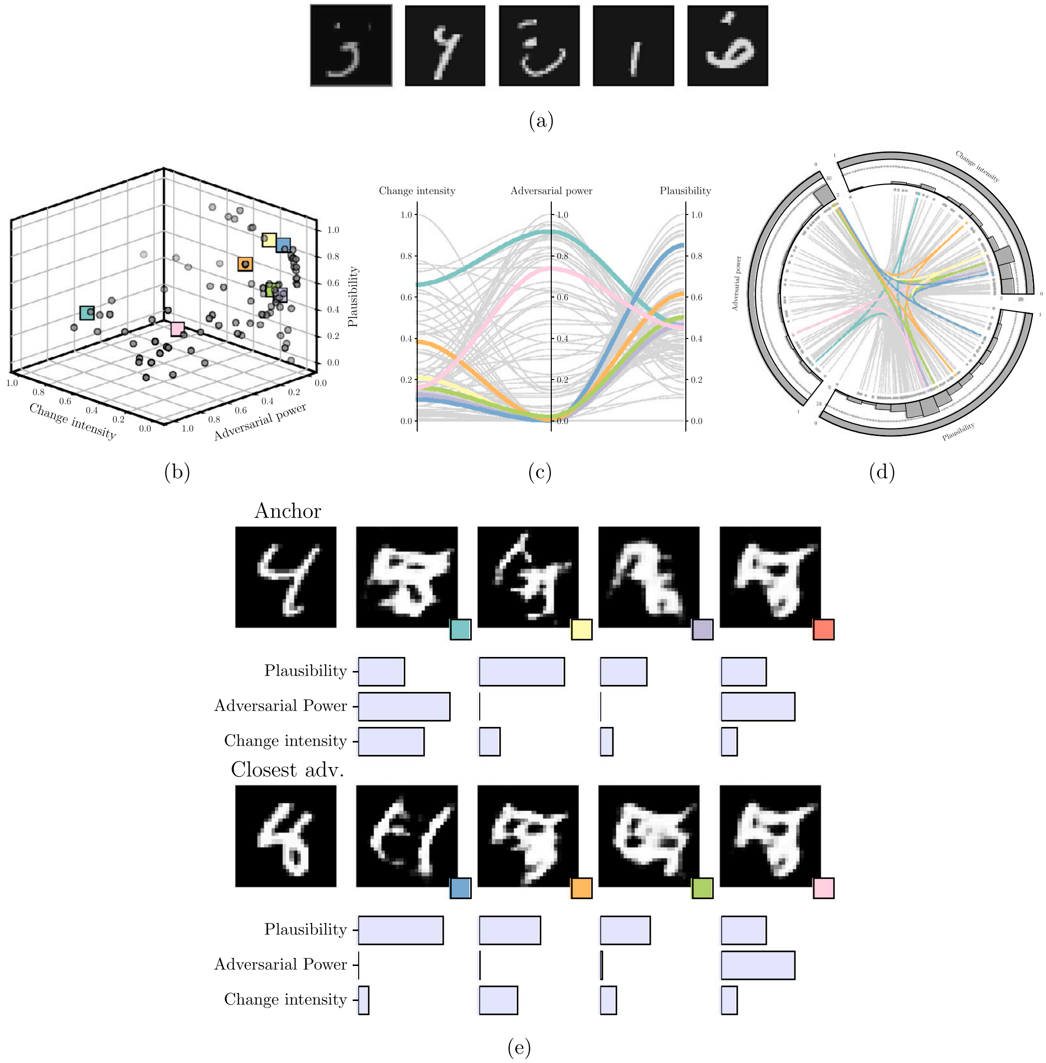


Fig. 14. (a) Sample of the unfinished digits generated for supplementing the MNIST dataset as an additional class. These digits are designed to cover the gap found in the initial phase of experiment #5, which showed a weakness of the MNIST dataset concerning non-finished digits; (b, c, d and e) output of the framework when auditing the same model trained over the augmented dataset. In this case, images of the produced counterfactual instances do not conform to what human thinking expects to be a digit.

available) improves the overall trustworthiness of the audience in the model's output through the inspection of its corresponding counterfactual explanation and the assessment of the induced changes.

However, advanced use of this explanatory interface should regard other aspects to respond to the *so much for how much?* question in counterfactual analysis. This is, in essence, the ultimate purpose of the framework proposed in this paper, as well as the motivation for future research aimed at easing the interpretation of counterfactual explanations issued by the framework in more complex problems, comprising larger input and/or output dimensionalities (e.g. video classification and multi-modal classification tasks). Improving the understandability of the counterfactuals as per the cognitive feedback of the audience will be also actively investigated, for which mechanisms will be devised to bring the human cognitive skills into the algorithmic loop of the counterfactual generation framework. Finally, we also plan to explore the application of modern generative models, including pre-trained large language models and stable diffusion models, to better represent the space of plausible counterfactual modifications to a given input. In particular, we will investigate whether their capabilities to diversify their synthesized outputs are superior to the herein-utilized GAN architectures.

Ethical approval

This work does not raise any ethical issues.

CRediT authorship contribution statement

All authors contributed to the Writing and Editing, JDS, NDR, FH, AH were additionally responsible for funding acquisition and supervising. JDS and AB were additionally responsible for methodology, study design, validation, formal analysis, data curation and visualization. All authors read and approved the final version of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work has received funding support from the Basque Government (*Eusko Jaurlaritza*) through the Consolidated Research Group MATHMODE (IT1256-22) and by the "Centro para el Desarrollo Tecnologico Industrial" (CDTI) and the European Union (AI4ES project, grant no. CER-20211030). Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554. Parts of this work have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826078 (Feature Cloud). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. N. Díaz-Rodríguez was supported by the Spanish Government Juan de la Cierva Incorporación contract (IJC2019-039152-I), the Google Research Scholar Programme 2021 and a Marie Skłodowska-Curie Actions (MSCA) Postdoctoral Fellowship with agreement ID: 101059332.

References

[1] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin, From "where" to "what": towards human-understandable explanations through concept relevance propagation, arXiv preprint, arXiv:2206.03208, 2022.

[2] M.M. Al Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, R.R. Yager, Deep learning approach for active classification of electrocardiogram signals, Inf. Sci. 345 (2016) 340–354.

[3] D. Alvarez-Melis, T.S. Jaakkola, Towards robust interpretability with self-explaining neural networks, arXiv:1806.07538, 2018.

[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS ONE 10 (7) (2015) e0130140.

[5] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, F. Herrera, Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.

[6] A. Benítez-Hidalgo, A.J. Nebro, J. García-Nieto, I. Oregi, J. Del Ser, jmetalpy: a python framework for multi-objective optimization with metaheuristics, Swarm Evol. Comput. 51 (2019) 100,598.

[7] L. Breiman, Statistical modelling: the two cultures (with comments and a rejoinder by the author), Stat. Sci. 16 (3) (2001) 199–231.

[8] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, arXiv:1809.11096, 2018.

[9] R.M. Byrne, Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning, in: Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), 2019, pp. 6276–6282.

[10] A.X. Chang, T.A. Funkhouser, L.J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, F. Yu, ShapeNet: an information-rich 3D model repository, arXiv preprint, arXiv:1512.03012, 2015.

[11] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks, in: IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839–847.

[12] Y. Chen, J. Wang, Y. Liu, Strategic classification with a light touch: learning classifiers that incentivize constructive adaptation, arXiv preprint, arXiv:2011.00355, 2020.

[13] Y.L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications, arXiv preprint, arXiv:2103.04244, 2021.

[14] C.A.C. Coello, G.B. Lamont, D.A. Van Veldhuizen, Evolutionary Algorithms for Solving Multi-Objective Problems, vol. 5, Springer, 2007.

[15] J. Crabbé, M. van der Schaar, Concept activation regions: a generalized framework for concept-based explanations, Adv. Neural Inf. Process. Syst. 35 (2022) 2590–2607.

[16] J. Donahue, P. Krähenbühl, T. Darrell, Adversarial feature learning, arXiv preprint, arXiv:1605.09782, 2016.

[17] J. Gawlikowski, C.R.N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al., A survey of uncertainty in deep neural networks, Artif. Intell. Rev. 56 (Suppl. 1) (2023) 1513–1589.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Adv. in Neural Information Processing Systems, 2014, pp. 2672–2680.

[19] W. Hasperué, The master algorithm: how the quest for the ultimate learning machine will remake our world, J. Comput. Sci. Technol. 15 (2) (2015) 157–158.

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[21] Z. He, W. Zuo, M. Kan, S. Shan, X. Chen, AttGAN: facial attribute editing by only changing what you want, IEEE Trans. Image Process. 28 (11) (2019) 5464–5478.

[22] A. Holzinger, K. Keiblinger, P. Holub, K. Zatloukal, H. Müller, AI for life: trends in artificial intelligence for biotechnology, New Biotechnol. 74 (1) (2023) 16–24, https://doi.org/10.1016/j.nbt.2023.02.001.

[23] A. Holzinger, H. Mueller, Toward human-ai interfaces to support explainability and causability in medical ai, IEEE Computer 54 (10) (2021) 78–86, https://doi.org/10.1109/MC.2021.3092610.

[24] A. Holzinger, A. Saranti, A. Angerschmid, B. Finzel, U. Schmid, H. Mueller, Toward human-level concept learning: pattern benchmarking for AI algorithms, Patterns 4 (7) (2023) 1–21.

[25] A. Holzinger, A. Saranti, A. Angerschmid, C.O. Retzlaff, A. Gronauer, V. Pejakovic, F. Medel-Jimenez, T. Krexner, C. Gollob, K. Stampfer, Digital transformation in smart farm and forest operations needs human-centered AI: challenges and future directions, Sensors 22 (8) (2022) 3043.

[26] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[27] A.H. Karimi, G. Barthe, B. Balle, I. Valera, Model-agnostic counterfactual explanations for consequential decisions, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 895–905.

[28] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[29] A. Kasirzadeh, A. Smart, The use and misuse of counterfactuals in ethical machine learning, in: ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 228–236.

[30] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, arXiv preprint, arXiv:1312.6114, 2013.

[31] Z.C. Lipton, The mythos of model interpretability, Queue 16 (3) (2018) 31–57.

[32] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: International Conference on Computer Vision (ICCV), 2015, pp. 3730–3738.

[33] C. Marx, F. Calmon, B. Ustun, Predictive multiplicity in classification, in: International Conference on Machine Learning, PMLR, 2020, pp. 6765–6774.

[34] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint, arXiv:1411.1784, 2014.

[35] M. Pawelczyk, K. Broelemann, G. Kasneci, On counterfactual explanations under predictive multiplicity, in: Conference on Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 809–818.

[36] J. Pearl, D. Mackenzie, The Book of Why, Basic Books, New York, 2018.

[37] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, P. Flach, FACE: feasible and actionable counterfactual explanations, in: AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 344–350.

[38] K. Rawal, E. Kamar, H. Lakkaraju, Can I still trust you?: understanding the impact of distribution shifts on algorithmic recourses, arXiv preprint, arXiv:2012.11788, 2020.

[39] N.J. Roese, Counterfactual thinking, Psychol. Bull. 121 (1) (1997) 133.

[40] A. Saranti, M. Hudec, E. Mináriková, Z. Takáč, U. Großschedl, C. Koch, B. Pfeifer, A. Angerschmid, A. Holzinger, Actionable explainable AI (AxAI): a practical example with aggregation functions for adaptive classification and textual explanations for interpretable machine learning, Mach. Learn. Knowl. Extr. 4 (4) (2022) 924–953.

[41] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts, Data Min. Knowl. Discov. (2023) 1–59.

[42] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps, arXiv preprint, arXiv:1312.6034, 2013.

[43] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11,974–12,001.

[44] B. Ustun, A. Spangher, Y. Liu, Actionable recourse in linear classification, in: Conference on Fairness, Accountability, and Transparency, 2019, pp. 10–19.

[45] N. Van Hoeck, P.D. Watson, A.K. Barbey, Cognitive neuroscience of human counterfactual reasoning, Front. Human Neurosci. 9 (2015) 420.

[46] S. Verma, V. Boonsanong, M. Hoang, K.E. Hines, J.P. Dickerson, C. Shah, Counterfactual explanations and algorithmic recourses for machine learning: a review, arXiv:2010.10596, 2020, pp. 1–23.

[47] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the gdpr, Harv. J. Law Technol. 31 (2017) 841.

[48] J. Wu, C. Zhang, X. Zhang, Z. Zhang, W.T. Freeman, J.B. Tenenbaum, Learning 3D shape priors for shape completion and reconstruction, in: European Conference on Computer Vision (ECCV), 2018, pp. 1–17.

[49] F. Yu, Y. Zhang, S. Song, A. Seff, J. Xiao, Lsun: construction of a large-scale image dataset using deep learning with humans in the loop, arXiv preprint, arXiv:1506.03365, 2015.

[50] J.Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A.A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, arXiv preprint, arXiv:1711.11586, 2017.