STEEX: Steering Counterfactual Explanations with Semantics

Paul Jacob[1], Éloi Zablocki[1], Hédi Ben-Younes[1], Mickaël Chen[1], Patrick Pérez[1], and Matthieu Cord[1,2]

[1] Valeo.ai
[2] Sorbonne University

Abstract. As deep learning models are increasingly used in safety-critical applications, explainability and trustworthiness become major concerns. For simple images, such as low-resolution face portraits, synthesizing visual counterfactual explanations has recently been proposed as a way to uncover the decision mechanisms of a trained classification model. In this work, we address the problem of producing counterfactual explanations for high-quality images and complex scenes. Leveraging recent semantic-to-image models, we propose a new generative counterfactual explanation framework that produces plausible and sparse modifications which preserve the overall scene structure. Furthermore, we introduce the concept of "region-targeted counterfactual explanations", and a corresponding framework, where users can guide the generation of counterfactuals by specifying a set of semantic regions of the query image the explanation must be about. Extensive experiments are conducted on challenging datasets including high-quality portraits (CelebAMask-HQ) and driving scenes (BDD100k). Code is available at: https://github.com/valeoai/STEEX

Keywords: Explainable AI, Counterfactual Analysis, Visual explanations, Region-targeted Counterfactual Explanation.

1 Introduction

Deep learning models are now used in a wide variety of application domains, including safety-critical ones. As the underlying mechanisms of these models remain opaque, explainability and trustworthiness have become major concerns. In computer vision, *post-hoc* explainability often amounts to producing saliency maps, which highlight regions on which the model grounded the most its decision [59,38,2,40,43,54,13]. While these explanations show *where* the regions of interest for the model are, they fail to indicate *what* specifically in these regions leads to the obtained output. A desirable explanation should not only be *region*-based but also *content*-based by expressing in some way how the content of a region influences the outcome of the model. For example, in autonomous driving, while it is useful to know that a stopped self-driving car attended the traffic light, it is paramount to know that the red color of the light was decisive in the process.

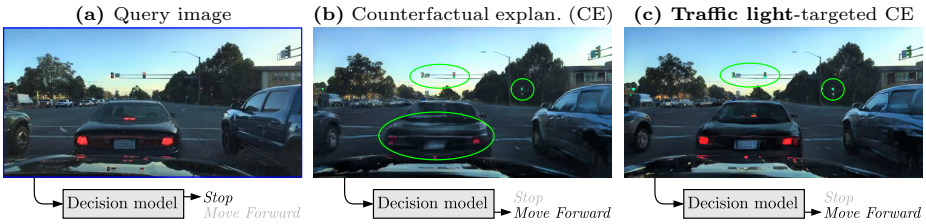

Fig. 1: Overview of counterfactual explanations generated by our framework STEEX. Given a trained model and a query image (a), a counterfactual explanation is an answer to the question "*What other image, slightly different and in a meaningful way, would change the model's outcome?*" In this example, the 'Decision model' is a binary classifier that predicts whether or not it is possible to move forward. On top of explaining decisions for large and complex images (b), we propose 'region-targeted counterfactual explanations' (c), where produced counterfactual explanations only target specified semantic regions. Green ellipses are manually provided to highlight details

In the context of simple tabular data, *counterfactual explanations* have recently been introduced to provide fine content-based insights on a model's decision [48,47,8]. Given an input query, a counterfactual explanation is a version of the input with *minimal* but *meaningful* modifications that change the output decision of the model. *Minimal* means that the new input must be as similar as possible to the query input, with only sparse changes or in the sense of some distance to be defined. *Meaningful* implies that changes must be semantic, i.e., human-interpretable. This way, a counterfactual explanation points out in an understandable way *what* is important for the decision of the model by presenting a close hypothetical reality that contradicts the observed decision. As they are contrastive and as they usually focus on a small number of feature changes, counterfactuals can increase user's trust in the model [39,56,53]. Moreover, these explanations can also be leveraged by machine learning engineers, as they can help to identify spurious correlations captured by a model [45,55,36]. Despite growing interest, producing visual counterfactual explanations for an image classification model is especially challenging as naively searching for small input changes results in adversarial perturbations [44,18,14,33,6]. To this date, there only exists a very limited number of counterfactual explanation methods able to deal with image classifiers [19,50,41,36]. Yet, these models present significant limitations, as they either require a target image of the counterfactual class [19,50] or can only deal with classification settings manipulating simple images such as low-resolution face portraits [41,36].

In this work, we tackle the generation of counterfactual explanations for deep classifiers operating on large images and/or visual scenes with complex structures. Dealing with such images comes with unique challenges, beyond technical issues. Indeed, because of scene complexity, it is likely that the model's decision can be changed by many admissible modifications in the input. For a driving action classifier, it could be for instance modifying the color of traffic lights, the road markings or the visibility conditions, but also adding new elements to

the scene such as pedestrians and traffic lights, or even replacing a car on the road with an obstacle. Even if it was feasible to provide an exhaustive list of counterfactual explanations, the task of selecting which ones in this large collection are relevant would fall on the end-user, hindering the usability of the method. To limit the space of possible explanations while preserving sufficient expressivity, we propose that the overall structure of the query image remains untouched when creating the counterfactual example. Accordingly, through semantic guidance, we impose that a generated counterfactual explanation respects the original layout of the query image.

Our model, called STEEX for STEering counterfactual EXplanations with semantics, leverages recent breakthroughs in semantic-to-real image synthesis [37,32,60]. A pre-trained encoder network decomposes the query image into a spatial layout structure and latent representations encoding the content of each semantic region. By carefully modifying the latent codes towards a different decision, STEEX is able to generate meaningful counterfactuals with relevant semantic changes and a preserved scene layout, as illustrated in Fig. 1b. Additionally, we introduce a new setting where users can guide the generation of counterfactuals by specifying which semantic region of the query image the explanation must be about. We coin "region-targeted counterfactual explanations" such generated explanations where only a subset of latent codes is allowed to be modified. In other words, such explanations are answers to questions such as *"How should the traffic lights change to switch the model's decision?"*, as illustrated in Fig. 1c. To validate our claims, extensive experiments of STEEX are conducted on a variety of image classification models trained for different tasks, including self-driving action decision on the BDD100k dataset, and high-quality face recognition networks trained on CelebAMask-HQ. Besides, we investigate how explanations for different decision models can hint at their distinct and specific behaviors.

To sum up, our contributions are as follows:

- We tackle the generation of visual counterfactual explanations for classifiers dealing with large and/or complex images.
- By leveraging recent semantic-to-image generative models, we propose a new framework capable of generating counterfactual explanations that preserve the semantic layout of the image.
- We introduce the concept of "region-targeted counterfactual explanations" to target specified semantic regions in the counterfactual generation process.
- We validate the quality, plausibility and proximity to their query, of obtained explanations with extensive experiments, including classification models for high-quality face portraits and complex urban scenes.

2 Related Work

The black-box nature of deep neural networks has led to the recent development of many explanation methods [3,16,1,12]. In particular, our work is grounded within the *post-hoc* explainability literature aiming at explaining a

trained model, which contrasts with approaches building interpretable models *by design* [57,9]. Post-hoc methods can be either be *global* if they seek to explain the model in its entirety, or *local* when they explain the prediction of the model for a specific instance. Global approaches include model translation techniques, that distill the black-box model into a more interpretable one [15,20], or the more recent disentanglement methods that search for latent dimensions of the input space that are, at the level of the dataset, correlated with the output variations of the target classifier [25,28]. Instead, in this paper, we focus on *local* methods that provide explanations, tailored to a given image.

Usually, *post-hoc local* explanations of vision models are given in the form of saliency maps, which attribute the output decision to image regions. Gradient-based approaches compute this attribution using the gradient of the output with respect to input pixels or intermediate layers [38,43,34,4]. Differently, perturbation-based approaches [54,58,13,49] evaluate how sensitive to input variations is the prediction. Other explainability methods include locally fitting a more interpretable model such as a linear function [35] or measuring the effect of including a feature with game theory tools [30]. However, these methods only provide information on *where* are the regions of interest for the model but do not tell *what* in these regions is responsible for the decision.

Counterfactual explanations [48], on the other hand, aim to inform a user on why a model M classifies a specific input x into class y instead of a *counter class* $y' \neq y$. To do so, a *counterfactual example* x' is constructed to be similar to x but classified as y' by M. Seminal methods have been developed in the context of low-dimensional input spaces, like the ones involved in credit scoring tasks [48]. Naive attempts to scale the concept to higher-dimensional input spaces, such as natural images, face the problem of producing adversarial examples [44,18,31,6], that is, *imperceptible* changes to the query image that switch the decision. While the two problems have similar formulations, their goals are in opposition [14,33] since counterfactual explanations must be understandable, achievable, and informative for a human. Initial attempts to counterfactual explanations of vision models would explain a decision by comparing the image x to one or several real instances classified as y' [21,19,50]. However, these discriminative counterfactuals do not produce natural images as explanations, and their interpretability is limited when many elements vary from one image to another.

To tackle these issues, generative methods leverage deep generative models to produce counterfactual explanations. For instance, DiVE [36] is built on β-TCVAE [11] and takes advantage of its disentangled latent space to discover such meaningful sparse modifications. With this method, it is also possible to generate multiple orthogonal changes that correspond to different valid counterfactual examples. Progressive Exaggeration (PE) [41], instead, relies on a Generative Adversarial Network (GAN) [17] conditioned on a perturbation value that is introduced as input in the generator via conditional batch normalization. PE modifies the query image so that the prediction of the decision model is shifted by this perturbation value towards the counter class. By applying this modification multiple times, and by showing the progression, PE highlights adjustments that

would change the decision model's output. Unfortunately, none of these previous works is designed to handle complex scenes. The β-TCVAE used in DiVE hardly scales beyond small centered images, requiring specifically-designed enhancement methods [27,42], and PE performs style-based manipulations that are unsuited images with multiple small independent objects of interest. Instead, our method relies on segmentation-to-image GANs [32,37,60], that have demonstrated good generative capabilities on high-quality images containing multiple objects.

3 Model STEEX

We now describe our method to obtain counterfactual explanations with semantic guidance. First, we formalize the generative approach for visual counterfactual explanations in Sec. 3.1. Within this framework, we then incorporate a semantic guidance constraint in Sec. 3.2. Next we propose in Sec. 3.3 a new setting where the generation targets specified semantic regions. Finally, Sec. 3.4 details the instantiation of each component. An overview of STEEX is presented in Fig. 2.

3.1 Visual Counterfactual Explanations

Consider a trained differentiable machine learning model M, which takes an image $x^I \in \mathcal{X}$ from an input space $\mathcal{X}$ and outputs a prediction $y^I = M(x^I) \in \mathcal{Y}$. A counterfactual explanation for the obtained decision y^I is an image x which is as close to the image x^I as possible, but such that $M(x) = y$ where $y \neq y^I$ is another class. This problem can be formalized and relaxed as follows:

$$\text{argmin}_{x \in \mathcal{X}} \, L_{\text{decision}}(M(x), y) + \lambda L_{\text{dist}}(x^I, x), \tag{1}$$

where L_{decision} is a classification loss, L_{dist} measures the distance between images, and the hyperparameter λ balances the contribution of the two terms.

In computer vision applications where input spaces are high-dimensional, additional precautions need to be taken to avoid ending up with adversarial examples [44,14,33,6]. To prevent those uninterpretable perturbations, which leave the data manifold by adding imperceptible high-frequency patterns, counterfactual methods impose that visual explanations lie in the original input domain $\mathcal{X}$. Incorporating this in-domain constraint can be achieved by using a deep generator network as an implicit prior [46,5]. Consider a generator $G : z \mapsto x$ that maps vectors z in latent space $\mathcal{Z}$ to in-distribution images x. Searching images only in the output space of such a generator would be sufficient to satisfy the in-domain constraint, and the problem now reads:

$$\text{argmin}_{z \in \mathcal{Z}} \, L_{\text{decision}}(M(G(z)), y) + \lambda L_{\text{dist}}(x^I, G(z)). \tag{2}$$

Eq. 2 formalizes practices introduced in prior works [36,41] that also aim to synthesize counterfactual explanations for images.

Furthermore, assuming that a latent code z^I exists and can be recovered for the image x^I, we can express the distance loss directly in the latent space $\mathcal{Z}$:

$$\text{argmin}_{z \in \mathcal{Z}} \, L_{\text{decision}}(M(G(z)), y) + \lambda L_{\text{dist}}(z^I, z). \tag{3}$$

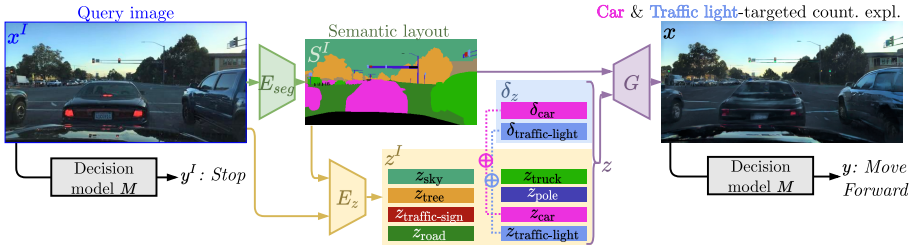

Fig. 2: Overview of STEEX. The query image x^I is first decomposed into a semantic map S^I and $z^I = (z_c)_{c=1}^N$, a collection of N semantic embeddings which encode each the aspect of their corresponding semantic category c. The perturbation δ_z is optimized such that the generated image $x = G(S^I, z^I + \delta_z)$ is classified as y by the decision model M, while staying small. As the generator uses the semantic layout S^I of the query image x^I, the generated counterfactual explanation x retains the original image structure. The figure specifically illustrates the region-targeted setting, where only the subset {'car', 'traffic light'} of the semantic style codes is targeted

By searching for an optimum in a low-dimensional latent space rather than in the raw pixel space, we operate over inputs that have a higher-level meaning, which is reflected in the resulting counterfactual examples.

3.2 Semantic-Guided Counterfactual Generation

The main objective of our model is to scale counterfactual image synthesis to large and complex scenes involving multiple objects within varied layouts. In such a setting, identifying and interpreting the modifications made to the query image is a hurdle to the usability of counterfactual methods. Therefore we propose to generate counterfactual examples that preserve the overall structure of the query and, accordingly, design a framework that optimizes under a fixed semantic layout. Introducing semantic masks for counterfactual explanations comes with additional advantages. First, we can leverage semantic-synthesis GANs that are particularly well-suited to generate diverse complex scenes [32,60,37]. Second, it provides more control over the counterfactual explanation we wish to synthesize, allowing us to target the changes to a specific set of semantic regions, as we detail in Sec. 3.3. To do so, we adapt the generator G and condition it on a semantic mask S that associates each pixel to a label indicating its semantic category (for instance, in the case of a driving scene, such labels can be cars, road, traffic signs, etc.). The output of the generator $G : (S, z) \mapsto x$ is now restricted to follow the layout indicated by S. We can then find a counterfactual example for image x^I that has an associated semantic mask S^I by optimizing the following objective:

$$\text{argmin}_{z \in \mathcal{Z}} L_{\text{decision}}(M(G(S^I, z)), y) + \lambda L_{\text{dist}}(z^I, z). \tag{4}$$

This formulation guarantees that the semantic mask S^I of the original scene is kept as is in the counterfactuals.

3.3 Region-Targeted Counterfactual Explanations

We introduce a new setting enabling finer control in the generation of counterfactuals. In this setup, a user specifies a set of semantic regions that the explanation must be about. For example, in Fig. 2, the user selects 'car' and 'traffic light', and the resulting counterfactual is only allowed to alter these regions. Such a selection allows studying the influence of different semantic concepts in the image for the target model's behavior. In practice, given a semantic mask S with N classes, we propose to decompose z into N vectors, $z = (z_c)_{c=1}^{N}$, where each z_c is a latent vector associated with one class in S. With such a formulation, it becomes possible to target a subset $C \subset \{1, \dots, N\}$ for the counterfactual explanation. Region-targeted counterfactuals only optimize on the specified components $(z_c)_{c \in C}$, and all other latent codes remain unmodified.

3.4 Instantiation of STEEX

We now present the modeling choices we make for each part of our framework.

Generator G. The generator G can be any of the recent segmentation-to-image GANs [32,60,37] that transform a latent code z and a segmentation layout S into an image x. As such generators typically allow for a different vector z_c to be used for each class in the semantic mask [60,37], the different semantic regions can be modified independently in the output image. This property enables STEEX to perform region-targeted counterfactual explanations as detailed in Sec. 3.3.

Obtaining the code z^I. To recover the latent code z^I from the image x^I, we exploit the fact that in aforementioned frameworks [60,37], the generator G can be trained jointly, in an auto-encoding pipeline, with an encoder E_z that maps an image x^I and its associated segmentation layout S^I into a latent code z^I. Such a property ensures that we can efficiently compute this image-to-latent mapping and that there is indeed a semantic code that corresponds to each image, leading to an accurate reconstruction in the first place.

Obtaining the mask S^I. As query images generally have no associated annotated segmentation masks S^I, these need to be inferred. To do so, we add a segmentation network E_{seg} in the pipeline: we first obtain the map $S^I = E_{\text{seg}}(x^I)$ and then use the encoder: $z^I = E_z(x^I, S^I)$, so STEEX is applicable to any image.

Loss functions. The decision loss L_{dist} ensures that the output image x is classified as y by the decision model M. It is thus set as the negative log-likelihood of the targeted counter class y for $M(G(z))$:

$$L_{\text{decision}}(M(G(z)), y) = -\mathcal{L}(M(G(z))|y). \tag{5}$$

The distance loss L_{dist} is the sum of squared L2 distance between each semantic component of z^I and z:

$$L_{\text{dist}}(z^I, z) = \sum_{c=1}^{N} \|z_c^I - z_c\|_2^2. \tag{6}$$

We stress that Eq. 4 is optimized on the code z only. All of the network parameters (G, E_z and E_{seg}) remain frozen.

4 Experiments

We detail in Sec. 4.1 our experimental protocol to evaluate different aspects of generated counterfactuals: the plausibility and perceptual quality (Sec. 4.2) as well as the proximity to query images (Sec. 4.3). We then present in Sec. 4.4 region-targeted counterfactual explanations. In Sec. 4.5, we use STEEX to explain different decision models for the same task, and show that produced explanations hint at the specificities of each model. Finally, we present an ablation study in Sec. 4.6. Our code and pretrained models will be made available.

4.1 Experimental Protocol

We evaluate our method on five decision models across three different datasets. We compare against two recently proposed visual counterfactual generation frameworks, Progressive Exaggeration (PE) [41] and DiVE [36], previously introduced in Sec. 2. We report scores directly from their paper when available (CelebA) and used the public and official implementation to evaluate them otherwise (CelebAMask-HQ and BDD100k). We now present each dataset and the associated experimental setup.

BDD100k [52]. The ability of STEEX to explain models handling complex visual scenes is evaluated on the driving scenes of BDD100k. Most images of this dataset contain diversely-positioned objects that can have fine relationships with each other, and small details in size can be crucial for the global understanding of the scene (e.g., traffic light colors). The decision model to be explained is a *Move Forward* vs. *Stop/Slow down* action classifier trained on BDD-OIA [51], a 20,000-scene extension of BDD100k annotated with binary attributes representing the high-level actions that are allowed in a given situation. The image resolution is 512×256. The segmentation model E_{seg} is a DeepLabV3 [10] trained on a subset of 10,000 images annotated with semantic masks that cover 20 classes (e.g., road, truck, car, tree, etc.). On the same set, the semantic encoder E_z and the generator G are jointly trained within a SEAN framework [60]. Counterfactual scores are computed on the validation set of BDD100k.

CelebAMask-HQ [26]. CelebAMask-HQ contains 30,000 high-quality face portraits with semantic segmentation annotation maps including 19 semantic classes (e.g., skin, mouth nose, etc.). The portraits are also annotated with identity and 40 binary attributes, allowing us to perform a quantitative evaluation for high-quality images. Decision models to be explained are two DenseNet121 [23] binary classifiers trained to respectively recognize *Smile* and *Young* attributes. To obtain semantic segmentation masks for the query images, we instantiate E_{seg} with a DeepLabV3 [10] pre-trained on the 28,000-image training split. On the same split, the semantic encoder E_z and generator G are jointly learned within a SEAN framework [60]. Counterfactual explanations are computed on the 2000-image validation set, with images rescaled to the resolution 256×256.

CelebA [29]. CelebA contains 200,000 face portraits, annotated with identity and 40 binary attributes, but of smaller resolution (128×128 after processing)

and of lower quality compared to CelebAMask-HQ. STEEX is designed to handle more complex and larger images, but we include this dataset for the sake of completeness as previous works [41,36] use it as their main benchmark. We report their score directly from their respective papers and align our experiment protocol with the one described in [36]. As in previous works, we explain two decision models: a *Smile* classifier and a *Young* classifier, both with DenseNet121 architecture [23]. We obtain E_{seg} with a DeepLabV3 [10] trained on CelebAMask-HQ images. Then, we jointly train the semantic encoder E_z and generator G with a SEAN architecture [60] on the training set of CelebA. Explanations are computed on the 19,868-image validation split of CelebA.

Optimization scheme. As M and G are differentiable, we optimize z using ADAM [24] with a learning rate $1 \cdot 10^{-2}$ for 100 steps with $\lambda = 0.3$. Hyperparameters have been found on the training splits of the datasets.

4.2 Quality of the Counterfactual Explanations

We first ensure that the success rate of STEEX, i.e., the fraction of explanations that are well classified into the counter class, is higher than 99.5% for all of the five tested classifiers. Then, as STEEX's counterfactuals must be realistic and informative, we evaluate their perceptual quality.

Similarly with previous works [41,36], we use the Fréchet Inception Distance (FID) [22] between all explanations and the set of query images, and report this metric in Tab. 1. For each classifier, STEEX outperforms the baselines by a large margin, meaning that our explanations are more realistic-looking, which verifies that they belong to the input domain of the decision model.

Generating realistic counterfactuals for classifiers that deal with large and complex images is difficult, as reflected by large FID discrepancies between CelebA, CelebAMask-HQ and BDD100k. Scaling the generation of counterfactual explanations from 128×128 (CelebA) to 256×256 (CelebAMask-HQ) face portraits is not trivial as a significant drop in performance can be observed for all models, especially for DiVE. Despite our best efforts to train DiVE on BDD100k, we were unable to obtain satisfying 512×256 explanations, as all reconstructions were nearly uniformly gray. As detailed in Sec. 2, VAE-based models are indeed

Table 1: Perceptual quality, measured with FID↓. Five attribute classifiers are explained, across three datasets. Results of PE and DiVE are reported from original papers on CelebA. For CelebAMask-HQ and BDD100k, their models are retrained using their code. DiVE does not converge on BDD100k

FID ↓	CelebA		CelebAM-HQ		BDD100k
	Smile	Young	Smile	Young	Move For.
PE [41]	35.8	53.4	52.4	60.7	141.6
DiVE [36]	29.4	33.8	107.0	107.5	—
STEEX	**10.2**	**11.8**	**21.9**	**26.8**	**58.8**

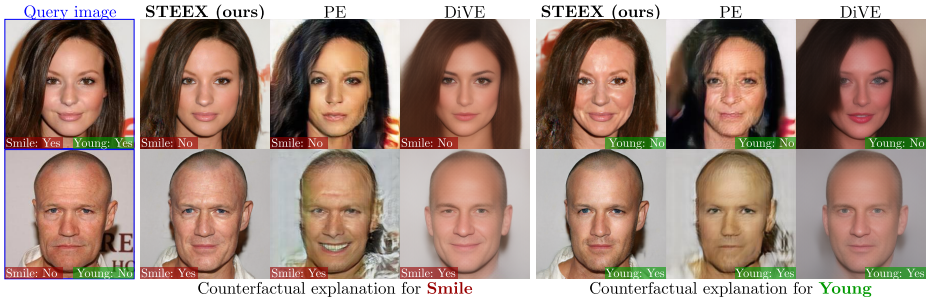

Fig. 3: Counterfactual explanations on CelebAMask-HQ, generated by STEEX (ours), PE, and DiVE. Explanations are generated for two binary classifiers, on *Smile* and *Young* attributes, at resolution 256 × 256. Other examples in the Supplementary

usually limited to images with a fairly regular structure, and they struggle to deal with the diversity of driving scenes.

We display examples of STEEX's counterfactual explanations on CelebAMask-HQ in Fig. 3, compared with PE [41] and DiVE [36]. For the *Smile* classifier, STEEX explains positive (top-row) and negative (bottom-row) smile predictions through sparse and photo-realistic modifications of the lips and the skin around the mouth and the eyes. Similarly, for the *Young* classifier, STEEX explain decisions by adding or removing facial wrinkles. In comparison, PE introduces high-frequency artifacts that harm the realism of generated examples. DiVE generates blurred images and applies large modifications so that it becomes difficult to identify the most crucial changes for the target model. Fig. 4 shows other samples for the action classifier on the BDD100k dataset, where we overlay green ellipses to point the reader's attention to significant region changes. STEEX finds sparse but highly semantic modifications to regions that strongly influence the output decision, such as the traffic light colors or the brake lights of a leading vehicle. Finally, the semantic guidance leads to a fine preservation of

Fig. 4: Counterfactual explanations on BDD100k. Explanations are generated for a binary classifier for the action *Move Forward*, with images at resolution 512 × 256. Our method finds interpretable, sparse and meaningful semantic modifications to the query image. Other examples are available in the Supplementary

the scene structure in STEEX's counterfactuals, achieving both global coherence and high visual quality.

4.3 Proximity to the Query Image

We now verify the *proximity* of counterfactuals to query images, as well as the *sparsity* of changes.

We first compare STEEX to previous work with respect to the **Face Verification Accuracy (FVA)**. The FVA is the percentage of explanations that preserve the person's identity, as revealed by a cosine similarity above 0.5 between features of the counterfactual and the query. Following previous works [41,36], features are computed by a pre-trained re-identification network on VGGFace2 [7]. As shown in Tab. 2, even if STEEX is designed for high-quality or complex scenes image classifiers, it reaches high FVA on the low-quality CelebA dataset. Moreover, STEEX significantly outperforms PE and DiVE on CelebAMask-HQ, showing its ability to scale up to higher image sizes. Again, DiVE suffers from the poor capacities of $\beta-$TCVAE to reconstruct high-quality images Sec. 2. To support this claim, we compute the FVA between query images and reconstructions with the $\beta-$TCVAE of DiVE and obtain 45.9%, which indicates a low reconstruction capacity.

We then measure the sparsity of explanations using the **Mean Number of Attributes Changed (MNAC)**. This metric averages the number of facial attributes that differ between the query image and its associated counterfactual explanation. As STEEX successfully switches the model's decision almost every time, explanations that obtain a low MNAC are likely to have altered only the necessary elements to build a counterfactual. Following previous work [36], we use an oracle ResNet pretrained on VGGFace2 [7], and fine-tuned on 40 attributes provided in CelebA/CelebAMask-HQ. As reported in Tab. 2, STEEX has a lower MNAC than PE and DiVE on both CelebA and CelebAMask-HQ. Conditioning the counterfactual generation on semantic masks helps obtaining small variations that are meaningful enough for the model to switch its decision. This property makes STEEX useful in practice and well-suited to explain image classifiers.

Table 2: Face Verification Accuracy (FVA↑) (%) and Mean Number of Attributes Changed (MNAC↓), on CelebA and CelebAMask-HQ. For PE and DiVE, CelebA scores come from the original papers, and we re-train their models using official implementations for CelebAMask-HQ

FVA ↑	CelebA		CelebAM-HQ	
	Smile	Young	Smile	Young
PE [41]	85.3	72.2	79.8	76.2
DiVE [36]	**97.3**	**98.2**	35.7	32.3
STEEX	96.9	97.5	**97.6**	**96.0**

MNAC ↓	CelebA		CelebAM-HQ	
	Smile	Young	Smile	Young
PE [41]	—	3.74	7.71	8.51
DiVE [36]	—	4.58	7.41	6.76
STEEX	**4.11**	**3.44**	**5.27**	**5.63**

Fig. 5: Semantic region-targeted counterfactual explanations on BDD100k. Explanations are generated for a binary classifier trained on the attribute *Move Forward*, at resolution 512×256. Each row shows explanations where we restrict the optimization process to one specific semantic region, on two examples: one where the model initially goes forward, and one where it initially stops. Significant modifications are highlighted within the green ellipses. Note that even when targeting specific regions, others may still slightly differ from the original image: this is mostly due to small errors in the reconstruction $G(S^I, z^I) \approx x^I$ (more details in the Supplementary)

4.4 Region-Targeted Counterfactual Explanations

As can be seen in Figs. 1b and 4, when the query image is complex, the counterfactual explanations can encompass multiple semantic concepts at the same time. In Fig. 1b for instance, in order to switch the decision of the model to *Move Forward*, the traffic light turns green and the car's brake lights turn off. It raises ambiguity about how these elements compound to produce the decision. In other words, "Are both changes necessary, or changing only one region is sufficient to switch the model's decision?".

To answer this question, we generate *region-targeted* counterfactual explanations, as explained in Sec. 3.3. In Fig. 1c, we observe that targeting the traffic light region can switch the decision of the model, despite the presence of a stopped car blocking the way. Thereby, region-targeted counterfactuals can help to identify potentially safety-critical issues with the decision model.

More generally, region-targeted counterfactual explanations empower the user to separately assess how different concepts impact the decision. We show in Fig. 5 qualitative examples of such region-targeted counterfactual explanations on the *Move Forward* classifier. On the one hand, we can verify that the decision model relies on cues such as the color of the traffic lights and brake lights of cars, as changing them often successfully switch the decision. On the other hand, we discover that changes in the appearance of buildings can flip the model's decision.

Fig. 6: Counterfactual explanations on CelebAMask-HQ for three different *Young* classifiers, namely M_{top}, M_{mid}, and M_{bot} that respectively only attend to the top, mid, and bottom parts of the image. Other examples are available in the Supplementary

Indeed, we see that green or red gleams on facades can fool the decision model into predicting *Move Forward* or *Stop* respectively, suggesting that the model could need further investigation before being safely deployed.

4.5 Analyzing Decision Models

An attractive promise of explainable AI is the possibility to detect and characterize biases or malfunctions of explained decision models. In this section, we investigate how specific are explanations to different decision models and if the explanations can point at the particularity of each model. In practice, we consider three decision models, namely M_{top}, M_{mid}, and M_{bot}, that were trained on images with masked out pixels except the for the top, middle, and bottom parts of the input respectively. Fig. 6 reports qualitative results, and we can identify that M_{top} has based its decisions mainly on the color of the hair, while M_{mid} uses the wrinkles on the face, and M_{bot} focuses on facial hair and the neck.

We also measure how much each semantic region has been modified to produce the counterfactual. Accordingly, we assess the impact of a semantic class c in the decision with the average value of $\|\delta_{z_c}\|_2 = \|z_c^I - z_c\|_2$ aggregated over the validation set. Note that while the absolute values of δ_{z_c} can be compared across the studied decision models, they cannot be directly compared across different semantic classes, as the z_c can be at different scales for different values of c in the generative model. To make this comparison in Tab. 3, we instead compute the value of δ_{z_c} for the target model *relatively* to the average value for all models.

Table 3: Most and least impactful semantic classes for a decision model relatively to others. The impact of a class for a given model has been determined as the average value of $\|\delta_{z_c}\|_2 = \|z_c^I - z_c\|_2$ for each semantic class c, relatively to the same value averaged for other models

Model	Most impactful	Least impactful
M_{top}	hat, hair, background	necklace, eyes, lips
M_{mid}	nose, glasses, eyes	necklace, neck, hat
M_{bot}	neck, necklace, cloth	eyes, brows, glasses

Table 4: Ablation study measuring the role of the distance loss L_{dist} in Eq. 4 and upper bound results that would be achieved with ground-truth segmentation masks

	Smile		Young	
	FID ↓	FVA ↑	FID ↓	FVA ↑
STEEX	21.9	97.6	26.8	96.0
without L_{dist}	29.7	65.2	45.7	37.0
with ground-truth segmentation	21.2	98.9	25.7	98.2

The semantic classes of most impact in Tab. 3 indicate how each decision model is biased towards a specific part of the face and ignores cues that are important for the other models.

4.6 Ablation Study

We propose an ablation study on CelebAMask-HQ, reported in Tab. 4, to assess the role of the distance loss L_{dist} and the use of predicted segmentation masks.

First, we evaluate turning off the distance loss by setting $\lambda = 0$, such that the latent codes z_c are no longer constrained to be close to z_c^I. Doing so, for both *Young* and *Smile* classifiers, the FVA and FID of STEEX degrade significantly, which respectively indicate that the explanation proximity to the real images is deteriorated and that the counterfactuals are less plausible. The distance loss is thus an essential component for STEEX.

Second, we investigate if the segmentation network E_{seg} is a bottleneck in STEEX. To do so, we replace the segmenter's outputs with ground-truth masks and generate counterfactual explanations with these. The fairly similar scores of both settings indicate that STEEX works well with inferred layouts.

5 Conclusion

In this work, we present STEEX, a method to generate counterfactual explanations for complex scenes, by steering the generative process using predicted semantics. To our knowledge, we provide the first framework for complex scenes where numerous elements can affect the decision of the target network. Experiments on driving scenes and high-quality portraits show the capacity of our method to finely explain deep classification models. For now, STEEX is designed to generate explanations that preserve the semantic structure. While we show the merits of this property, future work can consider how, within our framework, to handle operations such as shifting, removing, or adding objects, while keeping the explanation simple to interpret. Finally, we hope that the setup we propose in Sec. 4.5, when comparing explanations for multiple decision models with known behaviors, can serve as a basis to measure the interpretability of an explanation method.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access (2018) 3
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one (2015) 1
3. Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., Parekh, J.: Flexible and context-specific AI explainability: A multidisciplinary approach. CoRR **abs/2003.07703** (2020) 3
4. Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Ackel, L.J., Muller, U., Yeres, P., Zieba, K.: Visualbackprop: Efficient visualization of cnns for autonomous driving. In: ICRA (2018) 4
5. Bora, A., Jalal, A., Price, E., Dimakis, A.G.: Compressed sensing using generative models. In: ICML (2017) 5
6. Browne, K., Swift, B.: Semantics and explanation: why counterfactual explanations produce adversarial examples in deep neural networks. CoRR **abs/2012.10076** (2020) 2, 4, 5
7. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: FG (2018) 11
8. Chang, C., Creager, E., Goldenberg, A., Duvenaud, D.: Explaining image classifiers by counterfactual generation. In: ICLR (2019) 2
9. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.: This looks like that: Deep learning for interpretable image recognition. In: NeurIPS (2019) 4
10. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017) 8, 9
11. Chen, R.T.Q., Li, X., Grosse, R., Duvenaud, D.: Isolating sources of disentanglement in variational autoencoders. In: NeurIPS (2018) 4
12. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (XAI): A survey. CoRR (2020) 3
13. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: ICCV (2017) 1, 4
14. Freiesleben, T.: Counterfactual explanations & adversarial examples - common grounds, essential differences, and potential transfers. CoRR **abs/2009.05487** (2020) 2, 4, 5
15. Frosst, N., Hinton, G.E.: Distilling a neural network into a soft decision tree. In: Workshop on Comprehensibility and Explanation in AI and ML @AI*IA (2017) 4
16. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: DSSA (2018) 3
17. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) NeurIPS (2014) 4
18. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015) 2, 4
19. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual visual explanations. In: ICML (2019) 2, 4
20. Harradon, M., Druce, J., Ruttenberg, B.E.: Causal learning and explanation of deep neural networks via autoencoded activations. CoRR (2018) 4

21. Hendricks, L.A., Hu, R., Darrell, T., Akata, Z.: Grounding visual explanations. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV (2018) 4
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017) 9
23. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017) 8, 9
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 9
25. Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W.T., Isola, P., Globerson, A., Irani, M., Mosseri, I.: Explaining in style: Training a GAN to explain a classifier in stylespace. In: ICCV (2021) 4
26. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: CVPR (2020) 8
27. Lee, W., Kim, D., Hong, S., Lee, H.: High-fidelity synthesis with disentangled representation. In: ECCV (2020) 5
28. Li, Z., Xu, C.: Discover the Unknown Biased Attribute of an Image Classifier. In: The IEEE International Conference on Computer Vision (ICCV) (2021) 4
29. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015) 8
30. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: NeurIPS (2017) 4
31. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: CVPR (2016) 4
32. Park, T., Liu, M., Wang, T., Zhu, J.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019) 3, 5, 6, 7
33. Pawelczyk, M., Joshi, S., Agarwal, C., Upadhyay, S., Lakkaraju, H.: On the connections between counterfactual explanations and adversarial examples. CoRR **abs/2106.09992** (2021) 2, 4, 5
34. Rebuffi, S., Fong, R., Ji, X., Vedaldi, A.: There and back again: Revisiting backpropagation saliency methods. In: CVPR (2020) 4
35. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: SIGKDD (2016) 4
36. Rodríguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I.H., Charlin, L., Vázquez, D.: Beyond trivial counterfactual explanations with diverse valuable explanations. In: ICCV (2021) 2, 4, 5, 8, 9, 10, 11
37. Schönfeld, E., Sushko, V., Zhang, D., Gall, J., Schiele, B., Khoreva, A.: You only need adversarial supervision for semantic image synthesis. In: ICLR (2021) 3, 5, 6, 7
38. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017) 1, 4
39. Shen, Y., Jiang, S., Chen, Y., Yang, E., Jin, X., Fan, Y., Campbell, K.D.: To explain or not to explain: A study on the necessity of explanations for autonomous vehicles. CoRR (2020) 2
40. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: ICML (2017) 1
41. Singla, S., Pollack, B., Chen, J., Batmanghelich, K.: Explanation by progressive exaggeration. In: ICLR (2020) 2, 4, 5, 8, 9, 10, 11

42. Srivastava, A., Bansal, Y., Ding, Y., Hurwitz, C., Xu, K., Egger, B., Sattigeri, P., Tenenbaum, J., Cox, D.D., Gutfreund, D.: Improving the reconstruction of disentangled representation learners via multi-stage modelling. CoRR **abs/2010.13187** (2020) 5
43. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML (2017) 1, 4
44. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: ICLR (2014) 2, 4, 5
45. Tian, Y., Pei, K., Jana, S., Ray, B.: Deeptest: automated testing of deep-neural-network-driven autonomous cars. In: ICSE (2018) 2
46. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Deep image prior. IJCV (2020) 5
47. Verma, S., Dickerson, J.P., Hines, K.: Counterfactual explanations for machine learning: A review. CoRR **abs/2010.10596** (2020) 2
48. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law & Technology (2017) 2, 4
49. Wagner, J., Köhler, J.M., Gindele, T., Hetzel, L., Wiedemer, J.T., Behnke, S.: Interpretable and fine-grained visual explanations for convolutional neural networks. In: CVPR (2019) 4
50. Wang, P., Vasconcelos, N.: SCOUT: self-aware discriminant counterfactual explanations. In: CVPR (2020) 2, 4
51. Xu, Y., Yang, X., Gong, L., Lin, H.C., Wu, T.Y., Li, Y., Vasconcelos, N.: Explainable object-induced action decision for autonomous vehicles. In: CVPR (2020) 8
52. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: CVPR (2020) 8
53. Zablocki, É., Ben-Younes, H., Pérez, P., Cord, M.: Explainability of vision-based autonomous driving systems: Review and challenges. CoRR **abs/2101.05307** (2021) 2
54. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014) 1, 4
55. Zhang, M., Zhang, Y., Zhang, L., Liu, C., Khurshid, S.: Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems. In: IEEE ASE (2018) 2
56. Zhang, Q., Yang, X.J., Robert, L.P.: Expectations and trust in automated vehicles. In: CHI (2020) 2
57. Zhang, Q., Wu, Y.N., Zhu, S.: Interpretable convolutional neural networks. In: CVPR (2018) 4
58. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. In: ICLR (2015) 4
59. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016) 1
60. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: SEAN: image synthesis with semantic region-adaptive normalization. In: CVPR (2020) 3, 5, 6, 7, 8, 9