STEEX: Steering Counterfactual Explanations with Semantics

Paul Jacob[1] Éloi Zablocki[1] Hédi Ben-Younes[1] Mickaël Chen[1] Patrick Pérez[1]
Matthieu Cord[1,2]
[1] Valeo.ai [2] Sorbonne Université

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).

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 very 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 [2, 13, 36, 38, 41, 52, 57]. 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.

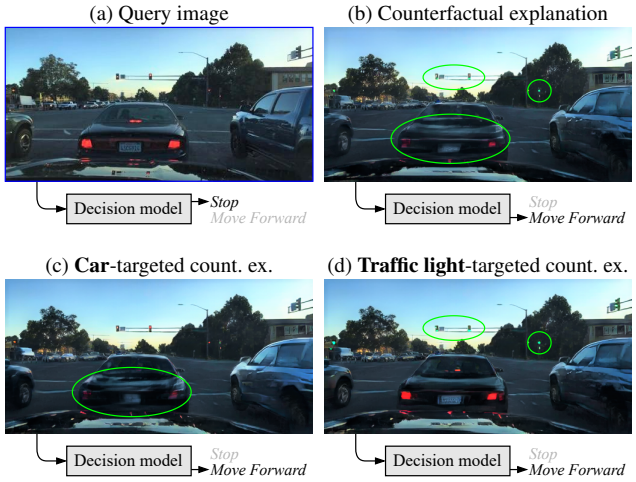

Figure 1. **Overview of counterfactual explanations generated by our framework STEEX.** Given a trained model and a query image, a counterfactual explanation is an answer to the question "*What other image, slightly different and in a meaningful way, would change the model's outcome?*" In this example, the 'Decision model' is a binary classifier that predicts whether or not it is possible to move forward. On top of handling large and complex images (top-right image), we propose 'region-targeted counterfactual explanations' (bottom images), 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 [8, 45, 46]. Given an input query, a counterfactual explanation is a version of the input with *minimal* but *meaningful* modifications that changes the output decision of the model. *Minimal* means that the new image must be as similar as possible to the query image, 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 [37, 51, 55]. Moreover, these explanations can also be leveraged by machine learning engineers, as they can help to identify spurious correlations captured by a model [34, 43, 53]. 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 [6, 14, 17, 31, 42]. To this date, there only exists a very limited number of counterfactual explanation methods able to deal with image classifiers [18, 34, 39, 48]. Yet, these models present significant limitations, as they either require a target image of the counterfactual class [18, 48] or can only deal with classification settings manipulating simple images such as low-resolution face portraits [34, 39].

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 [30, 35, 58]. 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 make the vehicle stop?"*, as illustrated in Figs. 1c and 1d. 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*.

To sum up, our contributions are as follows:

- We tackle the generation of 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" that enables 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 [1, 3, 12, 15]. 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* [9,54]. Usually, *post-hoc* 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 [4, 32, 36, 41]. Differently, perturbation-based approaches [13, 47, 52, 56] evaluate how sensitive to input variations is the prediction. Other explainability methods include locally fitting a more interpretable model such as a linear function [33] or measuring the effect of including a feature with game theory tools [28]. 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. Counterfactual explanations [46] 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, it constructs a *counterfactual example* x' that is similar to x but classified as y' by M. These methods have been developed in the context of low-dimensional input spaces, like the ones involved in credit scoring tasks [46]. Naive attempts to scale the concept to higher-dimensional input spaces, such as natural images, face the problem of producing adversarial examples [6, 17, 29, 42], that is, *im-*

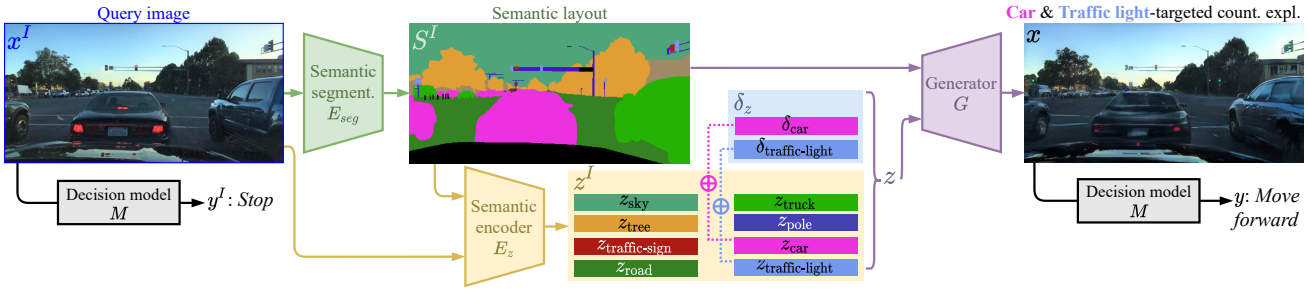

Figure 2. **Overview of STEEX.** The query image x^I is first decomposed into a semantic map S^I and $z^I = \{z_c\}_{c=1..N}$, a collection of N semantic embeddings which encode each the aspect of their corresponding semantic category c. The perturbation δ_z is optimized such that the generated image $x = G(S^I, z^I + \delta_z)$ 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.

perceptible changes to the query image that switch the decision. While the two problems have similar formulations, their goals are in opposition [14, 31] 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' [18,19,48]. 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.

Generative counterfactual explanations. On the other hand, generative methods leverage deep generative models to produce counterfactual examples that are similar to the query image but with some sparse changes in high-level attributes that switch the decision of the model. For instance, DiVE [34] 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) [39], instead, relies on a Generative Adversarial Network (GAN) [16] 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. Very recently, StylEx [24] trains a variant of StyleGAN2 [22] to obtain a classifier-specific disentangled style space. Then, a mining procedure is designed to extract a small set of meaningful attributes whose variation affects the classifier prediction. In both the case of PE and StyleEx, the image generation network is trained for a specific decision network, which makes it less flexible in real-world contexts. Moreover, none of those previous works are designed to handle complex scenes. DiVE relies on β-TCVAE, an unsupervised disentanglement method that hardly scales beyond small centered images, requiring specifically-designed enhancement methods [26,40]. As for PE and StylEx, they perform style-based manipulations that are not well-suited to deal with images that have multiple small independent objects of interest. Instead, our method relies on segmentation-to-image GANs [30, 35, 58], that have demonstrated good generative capabilities on high-quality images containing multiple objects.

3. Model STEEX

We now describe our method for obtaining 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, in Sec. 3.3, we propose 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:

$$\arg\min_{x \in \mathcal{X}} L_{\text{decision}}(M(x), y) + \lambda L_{\text{dist}}(x^I, x), \qquad (1)$$

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 [6, 14, 31, 42]. To prevent those uninterpretable perturbations, which leave the data manifold by adding high-frequency imperceptible 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 [5, 44]. Consider a generator $G : z \mapsto x$ that maps vectors z in latent space $\mathcal{Z}$ to in-distribution images. Searching only in the output space of such a generator would be sufficient to satisfy the in-domain constraint, and the problem now reads:

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

Eq. 2 formalizes practices introduced in prior works [34, 39] 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 $\mathcal{Z}$:

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

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 [30, 35, 58]. 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:

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

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 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, $z = (z_c)_{1 \leq c \leq N}$, 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 $z_c \in \{z_c | c \in C\}$, and all other latent codes remain unmodified.

3.4. Instantiation for STEEX

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

Generator G. The generator G can be instantiated with any of the recent Segmentation-to-Image GANs [30, 35, 58] 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 [35, 58], 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 [35, 58], 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} into the pipeline: we first obtain the map

$S^I = E_{seg}(x^I)$ and then use the encoder: $z^I = E_z(x^I, S^I)$. This makes STEEX 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 $M(G(z))$:

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

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

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

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 first detail our experimental protocol (Sec. 4.1) to evaluate different aspects of generated counterfactuals, namely the plausibility and perceptual quality (Sec. 4.2) as well as the proximity to query images (Sec. 4.3). Then, in Sec. 4.4, we present region-targeted counterfactual explanations. Finally, we present an ablation study in Sec. 4.5. 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) [39] and DiVE [34], 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 [50]. 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* [49], 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 [58]. Counterfactual scores are computed on the validation set of *BDD100k*.

CelebAMask-HQ [25]. *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 [21] 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 [58]. Counterfactual explanations are computed on the 2000-image validation set, with images rescaled to the resolution 256×256.

CelebA [27]. *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 [34, 39] 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 [34]. As in previous works, we explain two decision models: a *Smile* classifier and a *Young* classifier, both with DenseNet121 architecture [21]. 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 [58] 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 [23] 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 [34, 39], we use the Fréchet Inception Distance (FID) [20] 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

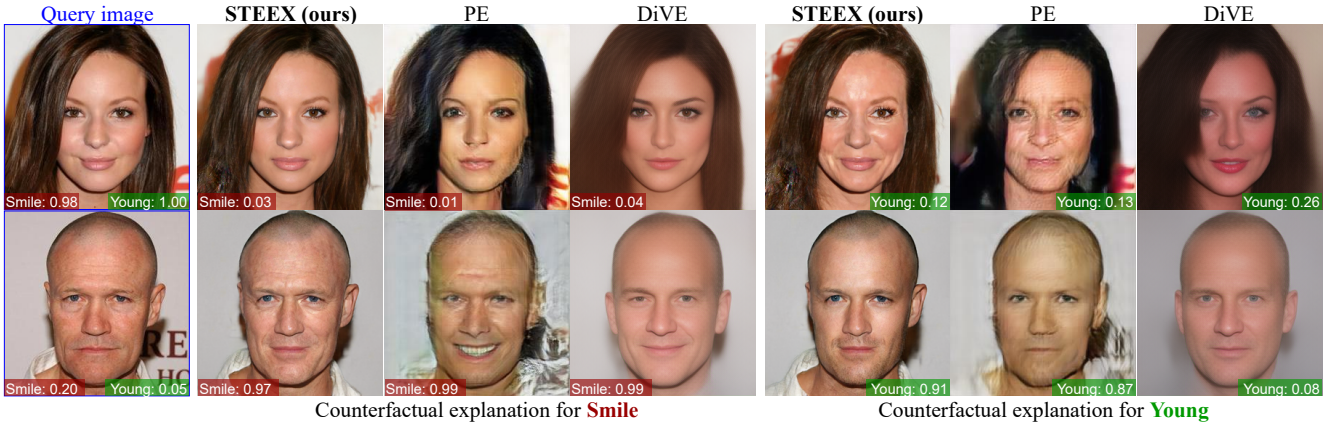

Figure 3. **Counterfactual explanations on *CelebAMask-HQ***, generated by STEEX, and the baselines PE and DiVE. Explanations are generated for two binary classifiers, on *Smile* and *Young* attributes, at resolution 256×256. Predicted scores are reported at the bottom of each image. Other examples are available in the Supplementary.

FID	*CelebA*		*CelebAM-HQ*		*BDD-100k*
	Smile	Young	Smile	Young	Move For.
PE [39]	35.8	53.4	52.4	60.7	141.6
DiVE [34]	29.4	33.8	107.0	107.5	—
STEEX	**10.2**	**11.8**	**21.9**	**26.8**	**58.8**

Table 1. **Explanation 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*, their models are retrained using their code.

more realistic-looking, hence more interpretable.

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 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 [39] and DiVE [34]. 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 the scene structure in STEEX's counterfactuals, achieving both global coherence and high visual quality.

4.3. Proximity of the counterfactual explanations

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 [34, 39], 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

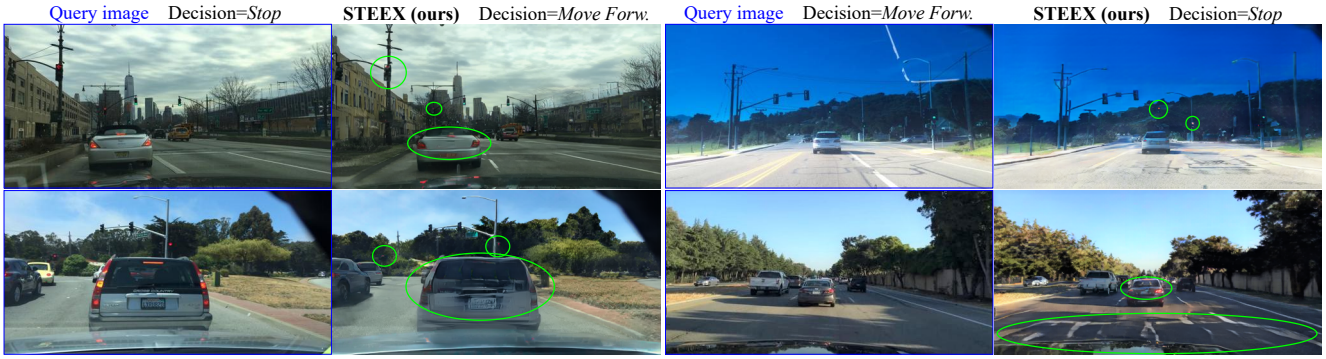

Figure 4. **Counterfactual explanations on *BDD100k*.** Explanations are generated for a binary classifier for the action *Move Forward*, with images at resolution 512×256. Our method finds interpretable, sparse and meaningful semantic modifications to the query image. Other examples are available in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

FVA	*CelebA*		*CelebAMask-HQ*	
	Smile	Young	Smile	Young
PE [39]	85.3	72.2	79.8	76.2
DiVE [34]	**97.3**	**98.2**	35.7	32.3
STEEX	96.9	97.5	**97.6**	**96.0**

Table 2. **Face Verification Accuracy (FVA) (%)**, 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*.

MNAC	*CelebA*		*CelebAMask-HQ*	
	Smile	Young	Smile	Young
PE [39]	—	3.74	7.71	8.51
DiVE [34]	—	4.58	7.41	6.76
STEEX	**4.11**	**3.44**	**5.27**	**5.63**

Table 3. **Mean Number of Attributes Changed (MNAC)**, on *CelebA* and *CelebAMask-HQ*. For PE and DiVE, results on *CelebA* are reported from the original papers (only available for *Young*) and we re-train models using official implementations to get scores on *CelebAMask-HQ*.

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 [34], we use an oracle ResNet pretrained on VGGFace2 [7], and fine-tuned on 40 attributes provided in *CelebA*/*CelebAMask-HQ*. As reported in Tab. 3, 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.

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 Figs. 1c and 1d, we observe that targeting independently, either the car region or the traffic light region, can switch the decision of the model, despite the presence of a red light or a stopped car blocking the way respectively. 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. 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.

Figure 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). To appreciate some of the fine details, please zoom in.

	Smile		Young	
	FID ↓	FVA ↑	FID ↓	FVA ↑
STEEX	21.9	97.6	26.8	96.0
wo/ L_{dist}	29.7	65.2	45.7	37.0
w/ g-t segm.	21.2	98.9	25.7	98.2

Table 4. **Ablation study** measuring the role of the distance loss L_{dist} in Eq. 4 and computation of upper bound results that would be achieved with ground-truth segmentation masks (g-t segm.).

4.5. 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 whether or not the segmentation network E_{seg} is a bottleneck in STEEX. To do so, we replace the segmenter's outputs with ground-truth segmentation masks and generate counterfactual explanations with these. The fairly similar scores in 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.

Limitations. STEEX is designed to generate explanations that preserve the semantic structure. While we show the merits of this property, it can sometimes be restrictive. For instance, operations such as shifting, removing, or adding objects, require optimizing the semantic layout, a problem that is left for future work.

Broader impacts. Our model provides new insights to analyze deep visual classifiers, to find limitations of visual models and build more fair, robust and trustworthy mod-

els. However, our approach is evaluated on face datasets. While they have the advantage of being extensively annotated, their use may raise concerns about privacy, and it is therefore paramount for our field to find better alternatives.

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 2018. 2

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 2015. 1

[3] Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché-Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi, and Jayneel Parekh. Flexible and context-specific AI explainability: A multidisciplinary approach. *CoRR*, abs/2003.07703, 2020. 2

[4] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry J Ackel, Urs Muller, Phil Yeres, and Karol Zieba. Visualbackprop: Efficient visualization of cnns for autonomous driving. In *ICRA*, 2018. 2

[5] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models. In *ICML*, 2017. 4

[6] Kieran Browne and Ben Swift. Semantics and explanation: why counterfactual explanations produce adversarial examples in deep neural networks. *CoRR*, abs/2012.10076, 2020. 2, 4

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018. 6, 7

[8] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *ICLR*, 2019. 1

[9] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan Su. This looks like that: Deep learning for interpretable image recognition. In *NeurIPS*, 2019. 2

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 5, 6

[11] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018. 3

[12] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *CoRR*, 2020. 2

[13] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017. 1, 2

[14] Timo Freiesleben. Counterfactual explanations & adversarial examples - common grounds, essential differences, and potential transfers. *CoRR*, abs/2009.05487, 2020. 2, 3, 4

[15] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *DSSA*, 2018. 2

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *NeurIPS*, 2014. 3

[17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 2

[18] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *ICML*, 2019. 2, 3

[19] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018. 3

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5

[21] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 5

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[24] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. Explaining in style: Training a GAN to explain a classifier in stylespace. In *ICCV*, 2021. 3

[25] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 5, 6

[26] Wonkwang Lee, Donggyun Kim, Seunghoon Hong, and Honglak Lee. High-fidelity synthesis with disentangled representation. In *ECCV*, 2020. 3

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5, 6

[28] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017. 2

[29] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *CVPR*, 2016. 2

[30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2, 3, 4

[31] Martin Pawelczyk, Shalmali Joshi, Chirag Agarwal, Sohini Upadhyay, and Himabindu Lakkaraju. On the connections between counterfactual explanations and adversarial examples. *CoRR*, abs/2106.09992, 2021. 2, 3, 4

[32] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *CVPR*, 2020. 2

[33] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *SIGKDD*, 2016. 2

[34] Pau Rodríguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam H. Laradji, Laurent Charlin, and David Vázquez. Beyond trivial counterfactual explanations with diverse valuable explanations. In *ICCV*, 2021. 2, 3, 4, 5, 6, 7

[35] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *ICLR*, 2021. 2, 3, 4

[36] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 1, 2

[37] Yuan Shen, Shanduojiao Jiang, Yanlin Chen, Eileen Yang, Xilun Jin, Yuliang Fan, and Katie Driggs Campbell. To explain or not to explain: A study on the necessity of explanations for autonomous vehicles. *CoRR*, 2020. 2

[38] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017. 1

[39] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. In *ICLR*, 2020. 2, 3, 4, 5, 6, 7, 8

[40] Akash Srivastava, Yamini Bansal, Yukun Ding, Cole Hurwitz, Kai Xu, Bernhard Egger, Prasanna Sattigeri, Josh Tenenbaum, David D Cox, and Dan Gutfreund. Improving the reconstruction of disentangled representation learners via multi-stage modelling. *CoRR*, abs/2010.13187, 2020. 3

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 1, 2

[42] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 2, 4

[43] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. Deeptest: automated testing of deep-neural-network-driven autonomous cars. In *ICSE*, 2018. 2

[44] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Deep image prior. *IJCV*, 2020. 4

[45] Sahil Verma, John P. Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *CoRR*, abs/2010.10596, 2020. 1

[46] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 2017. 1, 2

[47] Jörg Wagner, Jan Mathias Köhler, Tobias Gindele, Leon Hetzel, Jakob Thaddäus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *CVPR*, 2019. 2

[48] Pei Wang and Nuno Vasconcelos. SCOUT: self-aware discriminant counterfactual explanations. In *CVPR*, 2020. 2, 3

[49] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *CVPR*, 2020. 5, 6

[50] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 5, 6

[51] Éloi Zablocki, Hedi Ben-Younes, Patrick Pérez, and Matthieu Cord. Explainability of vision-based autonomous driving systems: Review and challenges. *CoRR*, abs/2101.05307, 2021. 2

[52] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1, 2

[53] Mengshi Zhang, Yuqun Zhang, Lingming Zhang, Cong Liu, and Sarfraz Khurshid. Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems. In *IEEE ASE*, 2018. 2

[54] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *CVPR*, 2018. 2

[55] Qiaoning Zhang, X. Jessie Yang, and Lionel Peter Robert. Expectations and trust in automated vehicles. In *CHI*, 2020. 2

[56] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *ICLR*, 2015. 2

[57] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1

[58] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. SEAN: image synthesis with semantic region-adaptive normalization. In *CVPR*, 2020. 2, 3, 4, 5, 6

STEEX: Steering Counterfactual Explanations with Semantics

— Supplementary Material —

A. Qualitative samples

In this section, we show additional samples of counterfactual explanations generated by STEEX, for the five classifiers mentioned in the main paper (trained on *CelebA*, *CelebAMask-HQ* and *BDD100k*).

STEEX on *CelebAMask-HQ*. In Fig. 7 and Fig. 8, we show samples for the *Smile*- and *Young*- classifiers on the *CelebAMask-HQ* dataset, with images of the size 256×256. The modifications found by STEEX to the images are plausible, understandable and easily traceable by a human due to their sparsity: they are mostly around the mouth for the *Smile*-classifier and on the skin and hair texture for the *Young*-classifier. Note that these explanations are *not* region-targeted, meaning that STEEX automatically selects the semantics to modify for the explanations.

STEEX on *CelebA*. In Fig. 6, we show samples for the *Smile*- and *Young*- classifiers on the *CelebA* dataset, with images of the size 128×128. STEEX applies both meaningful and sparse modifications to the query images and we can make similar observations as for *CelebAMask-HQ*.

Region-targeted counterfactuals on *CelebAMask-HQ*. In Fig. 9, we report examples of region-targeted counterfactual explanations on *CelebAMask-HQ*, for a binary classifier on the attribute *Young*. While the counterfactual explanations targeting the skin regions part mostly adds wrinkles to the faces, explanations on the hairy parts (hair and eyebrows) slightly turn them to gray. As skin-targeted counterfactuals are more convincing than hair-targeted counterfactuals, it may indicate that the decision model mostly relies on the skin texture and wrinkles to perform its '*Young*' classification.

STEEX on *BDD100k*. In Fig. 10 and Fig. 11, we show samples for the Move-forward classifier on the *BDD100k* dataset, with images of size 512×256. To explain '*Stop*' decisions, by providing counterfactual images where the decision model predicts '*Move forward*', several modifications can be observed depending on the image at hand, as reported by Fig. 10. The red light of traffic-lights can fade away (no light at all), or a green light can appear (top image). Besides, the back brake lights of the front vehicle can fade away as well. Interestingly, we observe on the top image that the brake lights of the front vehicle are more impacted than the brake light of the vehicle on the side. This may indicate that the decision model learned to mostly rely on the back lights of the front vehicle and not so much on vehicles of other lanes. On the other hand, in Fig. 11, to explain '*Move Forward*' decisions, by providing counterfactual images where the decision model predicts '*Stop*', modifications include green lights fading away, and rear brake lights of front cars turning on, as well as slight modification of the road texture which may indicate some spurious correlations learned by the decision model.

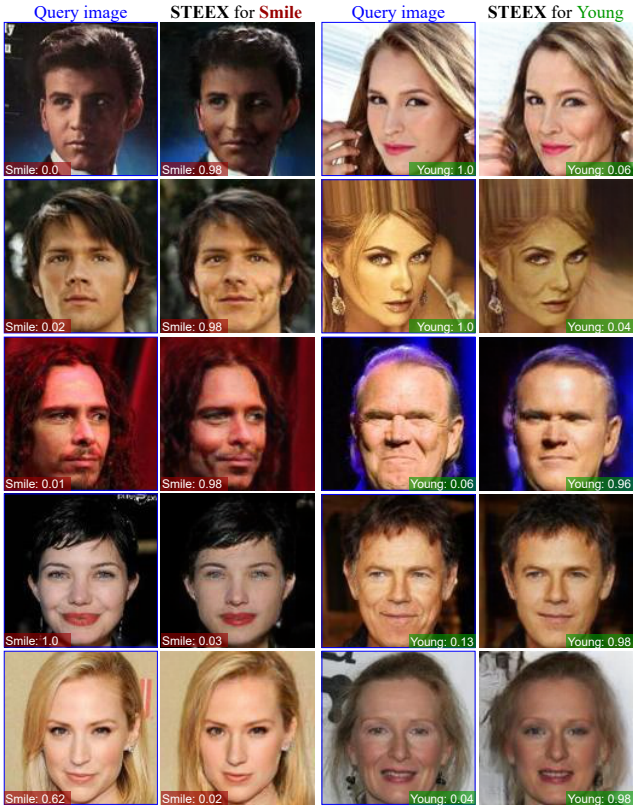

Figure 6. **Counterfactual explanations on *CelebA* generated by STEEX**. Explanations are generated for two binary classifiers, on *Smile* and *Young* attributes, at resolution 128×128. Predicted scores are reported at the bottom of each image.

STEEX vs. PE on *BDD100k*. In Fig. 12, we present a comparison between STEEX and PE [39] counterfactuals on the same query image for the *Move forward* classifier on *BDD100k* query images. We observe that counterfactual explanations produced by PE are blurred and, critically, they lose important details of the query image. On the other hand, STEEX successfully retrieves the details of the query image while applying plausible meaningful modifications. As explained in the main paper, we recall that, despite our

Figure 7. **Counterfactual explanations and reconstructions on *CelebAMask-HQ* generated by STEEX**. Explanations are generated for two binary classifiers, on *Smile* and *Young* attributes, at resolution 256×256. Predicted scores are reported at the bottom of each image.

best efforts, the adaptation of DiVE [34] to the driving scene dataset *BDD100k* produces mostly grey images. Indeed, DiVE suffers from the poor capacities of $\beta-$TCVAE to reconstruct high-quality images.

B. Reconstruction quality

In this section, we evaluate the impact of the *reconstruction* on the quality and sparsity of the generated counterfactuals. More precisely, we call '*reconstruction*' the image $G(S^I, z^I)$ generated from the predicted semantic mask

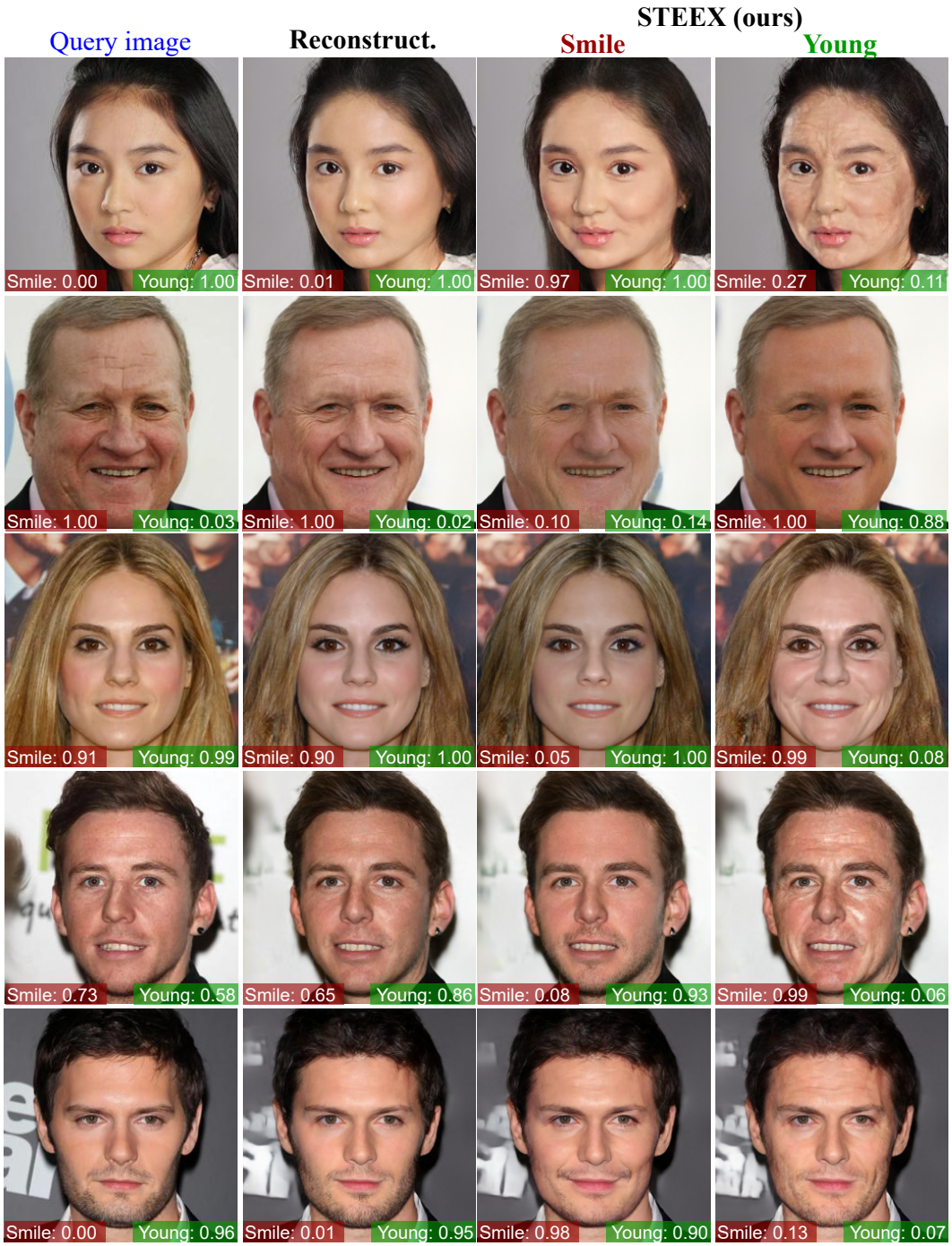

Figure 8. **Counterfactual explanations and reconstructions on *CelebAMask-HQ* generated by STEEX**. Explanations are generated for two binary classifiers, on *Smile* and *Young* attributes, at resolution 256×256. Predicted scores are reported at the bottom of each image.

$S^I = E_{seg}(x^I)$ and the semantic code $z^I = E_z(x^I, S^I)$ obtained on the query image x^I. Ensuring a good reconstruction quality is crucial. Indeed, the reconstructed image is the starting point of the optimization towards the counterfactual explanation. Thus, the reconstructed image must preserve as much as possible the content of the original query image. In a way, the quality of the reconstruction gives an upper bound to the quality of the generated counterfactual explanations.

In Tab. 5, we present a quantitative evaluation of the

Figure 9. **Region-targeted counterfactual explanations generated by STEEX on *CelebAMask-HQ***. Explanations are generated for a binary classifier on the *Young* attribute. From left to right: query images, counterfactual explanations on the skin, neck and nose, and counterfactual explanations on the hair and eyebrows. On the first set of explanations, STEEX mostly adds wrinkles, while on the second set, it greys slightly the hair.

quality (FID) and proximity (FVA, MNAC) between the reconstructed images $G(S^I, z^I)$ and the original query images x^I, for the three validation datasets. We recall that the reconstruction does not depend on the decision model M, but only on the pretrained networks E_{seg}, E_z, and G, which are dataset-specific. In each case, the results are close to the ones reported in Tab. 1, Tab. 2, and Tab. 3, meaning that the three metrics computed on our counterfactual explanations almost reach the proxy upper bounds. We can safely argue that our optimization process does not significantly degrade the images, both in terms of perceptual quality and proximity to the image query. Yet, improving the reconstruction quality, with better pretrained networks E_{seg}, E_z and G is thus an avenue for a quantitative boost in the results.

In Fig. 7, Fig. 8, Fig. 10 and Fig. 11, we show some examples of reconstructions obtained by STEEX on *CelebAMask-HQ* and *BDD100k*. Overall, a reconstructed image are highly faithful to its query image. However, looking at some close details, we can remark small changes between the query image and its reconstruction from semantics. This slight information loss then propagates on the final counterfactual explanations. Enhancing the reconstruction quality would yield more closeness between the query image and the counterfactual explanation.

	FID↓	MNAC↓	FVA↑ (%)
CelebA	8.4	2.04	99.3
CelebAMask-HQ	21.7	3.72	99.8
BDD100k	56.3	—	—

Table 5. **Evaluation of the reconstruction quality.** The reconstructed images are obtained with $G(S^I, z^I)$ and their quality is evaluated w.r.t. the original query images x^I with FID, MNAC and FVA metrics, for the three datasets used in this paper.

C. Technical details and code

C.1. Pseudo-code

In Alg. 1, we present the pseudo-code to generate a counterfactual explanation for the query image x^I on the model M with our method STEEX. It assumes that the semantic encoder E_z, the semantic segmentation network E_{seg} and the generator G have been previously pre-trained. The variable C is used to specify semantic regions in the region-targeted setting. In the general setting, the variable C simply includes all regions of the image.

C.2. Selection of the hyper-parameter λ

The hyper-parameter λ, which balances the respective contributions between the decision loss L_{decision} and the distance loss L_{dist}, was selected as the highest value such that the success-rate was almost perfect ($> 99.5\%$) on the training set of each dataset. For each of the five decision models, $\lambda = 0.3$. With higher values for λ, the decision is not always flipped. On the other hand, lower values imply that the obtained counterfactual explanation is further from the original query image and the person identity may be lost or more attributes may change. Setting $\lambda = 0$ implies that the distance loss has no contribution in the optimization, meaning that the only objective is the target decision.

We illustrate this in Fig. 13, where we show qualitative results with varying λ values. As a lower value for λ allows STEEX to find examples that are more distant to the query image, one can visualize the traits being more and more distorted towards the target decision, in a similar way to the method developed in Progressive Exaggeration [39]. With $\lambda = 0$, *i.e.* there is no distance penalty on the generated counterfactuals, images move away from the distribution of natural images, and we cannot consider that they are close enough to the type of images that the decision model M has been trained on, thus loosing the interest of the explanation. Still, it gives insights into the decision mode as it exaggerates important features for the decision model M.

Figure 10. **Counterfactual explanations on *BDD100k* generated by STEEX, where the decision model initially predicts '*Stop*'.** Explanations are generated for a binary classifier trained with the *BDD-OIA* dataset extension annotated with the attribute *Move forward*. The image resolution is 512×256.

C.3. Code

To reproduce our experiments, our code is publicly available at https://github.com/valeoai/STEEX.

C.4. Licenses

BDD100k data [50]. https://doc.bdd100k.com/license.html

BDD100k code. BSD 3-Clause License

BDD-OIA data [49]. No license provided

BDD-OIA code. BSD 3-Clause License

CelebA [27]. Agreement to use data on https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

CelebAMask-HQ [25]. Agreement to use data on https://mmlab.ie.cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html

SEAN [58] code. Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International https://github.com/ZPdesu/SEAN/blob/master/LICENSE.md

DiVE [34] code. Apache License 2.0 https://github.com/ElementAI/beyond-trivial-explanations/blob/master/LICENSE

PE [39] code. MIT Licence https://github.com/batmanlab/Explanation_by_Progressive_Exaggeration/blob/master/LICENSE.txt

DeepLabV3 [10] code. BSD 3-Clause License

Pytorch. BSD https://github.com/pytorch/pytorch/blob/master/LICENSE

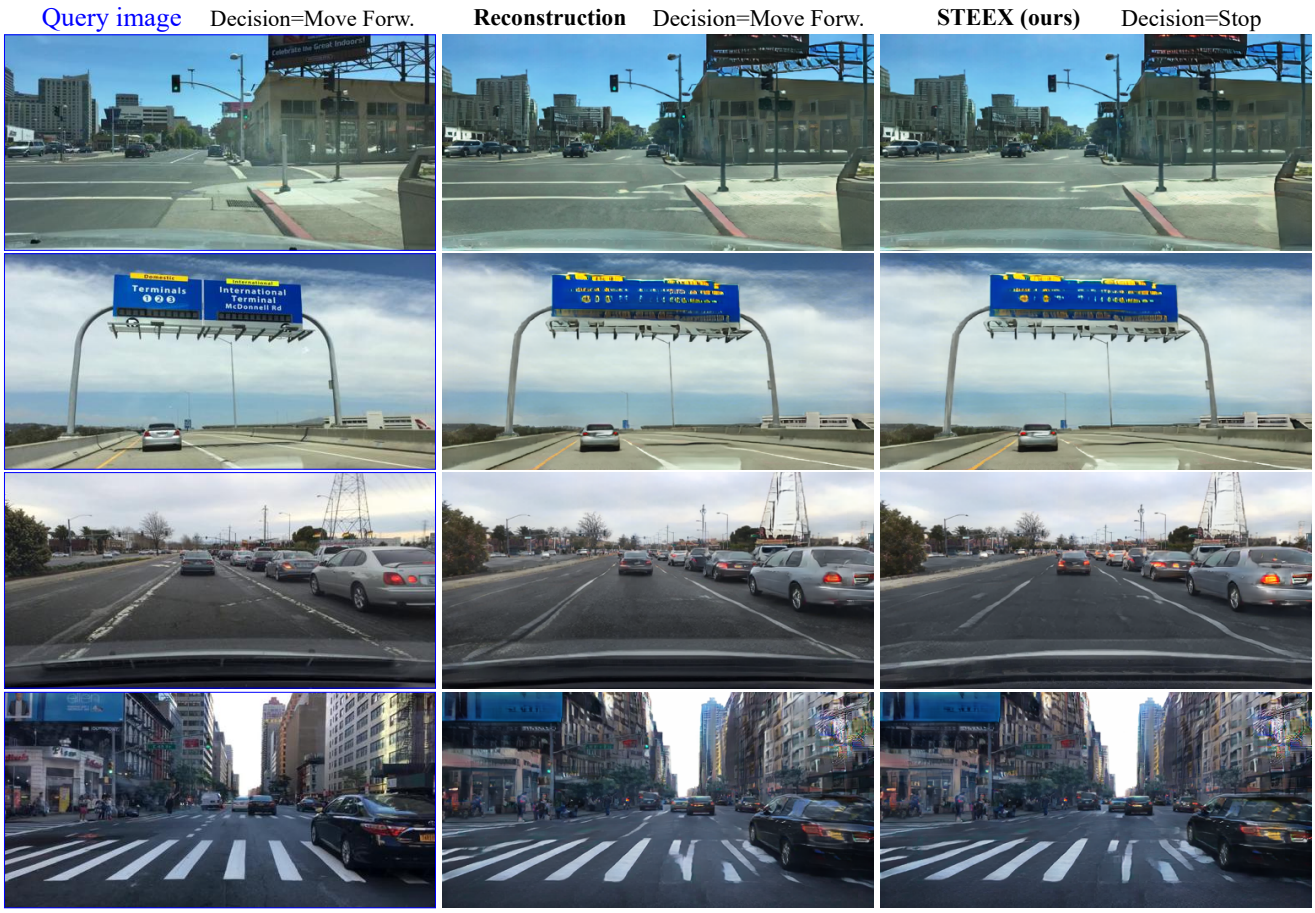

Figure 11. **Counterfactual explanations on *BDD100k* generated by STEEX, where the decision model initially predicts the '*Move forward*' class**. Explanations are generated for a binary classifier trained with the *BDD-OIA* dataset extension annotated with the attribute *Move forward*. The image resolution is 512×256.

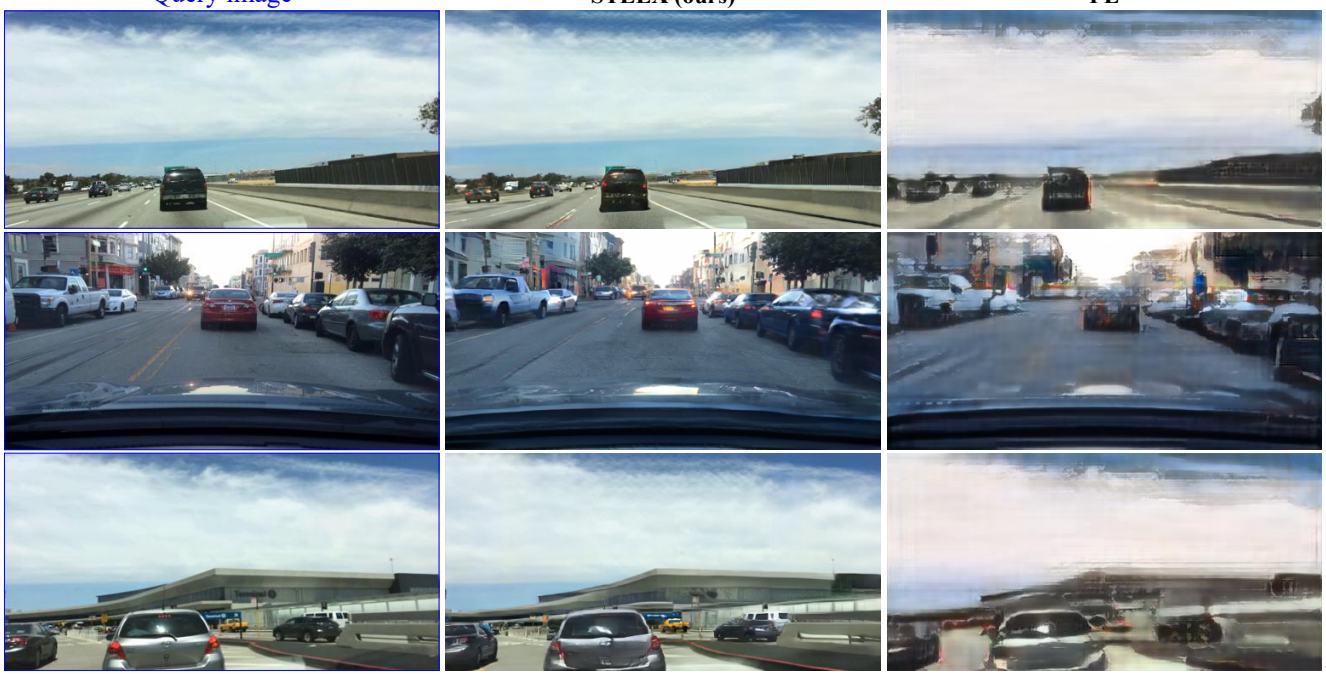

Figure 12. **Counterfactual explanations on *BDD100k* generated by STEEX compared to explanations generated by Progressive Exaggeration (PE) [39].** All images have a 512×256 resolution.

Algorithm 1 Pseudo-code for the counterfactual generation by STEEX. x^I is the query image and M is the binary decision model. C is the subset of regions to be targeted in the region-targeted setting. In the general setting, where counterfactual generation can modify the whole image, C simply includes all semantic regions. E_{seg} is a pretrained segmentation network, E_z is a pretrained latent encoder network, G is the generator network. The hyper-parameter λ balances the contribution between the two loss terms. N is the number of optimization steps. l_r is the learning rate for the optimization.

procedure GENERATE COUNTERFACTUAL(x^I, M, C, E_{seg}, E_z, G)
  $y^I \leftarrow M(x^I)$ ▷ Compute the original decision obtained for the query image
  if $y^I > 0.5$ **then** ▷ Get the target counter class y for the counterfactual explanation
    $y \leftarrow 0$
  else
    $y \leftarrow 1$
  end if
  $S^I \leftarrow E_{seg}(x^I)$ ▷ Compute the semantic layout of x^I
  $z^I \leftarrow E_z(x^I, S^I)$ ▷ Compute the latent codes for each semantic region
  $z \leftarrow z^I$ ▷ Initialize the latent code of the counterfactual explanation with z^I
  for $i \leftarrow 1, N$ **do** ▷ Make N optimization steps
    $x \leftarrow G(z, S^I)$ ▷ Generate x from the current code z, along with S^I
    $\tilde{y} \leftarrow M(x)$ ▷ Compute the model decision on x
    $L \leftarrow \mathcal{L}(\tilde{y}, y) + \lambda \sum_{c \in C} \|z_c^I - z_c\|_2^2$ ▷ Compute global objective inc. cross-entropy $\mathcal{L}$ and a distance penalty on z.
    $z \leftarrow \texttt{ADAM}(z, L, C, l_r)$ ▷ Update the code z with one gradient step, only on codes z_c with $c \in C$
  end for
  $x \leftarrow G(z, S^I)$ ▷ Compute the final counterfactual explanation
  return x
end procedure

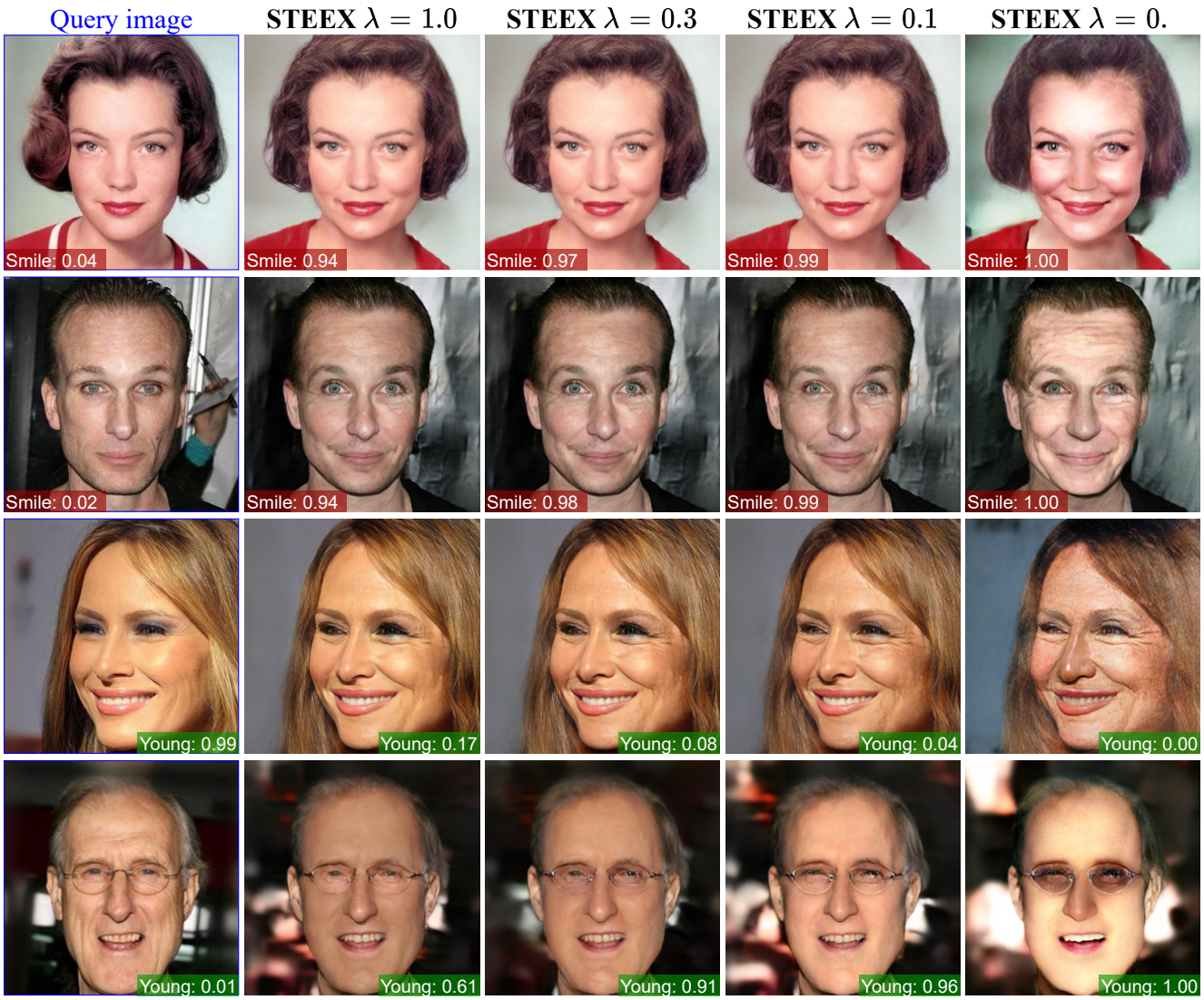

Figure 13. **Counterfactual explanations with various λ generated by STEEX**. The λ parameter balances the contribution of the loss $\mathcal{L}_{\text{dist}}$ with respect to the one of $\mathcal{L}_{\text{decision}}$. When λ is high, the decision is 'lightly' changed and the counterfactual explanation remains close to the query image. On the contrary, when λ is closer to zero, the generated counterfactual explanation is further from the query image and the decision is 'heavily' flipped.