STEEX: Steering Counterfactual Explanations with Semantics — Supplementary material —

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

A Additional 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. 1 and Fig. 2, 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. 3, 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. 4, 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 add 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. 5 and Fig. 6, 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

Fig. 1: 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.

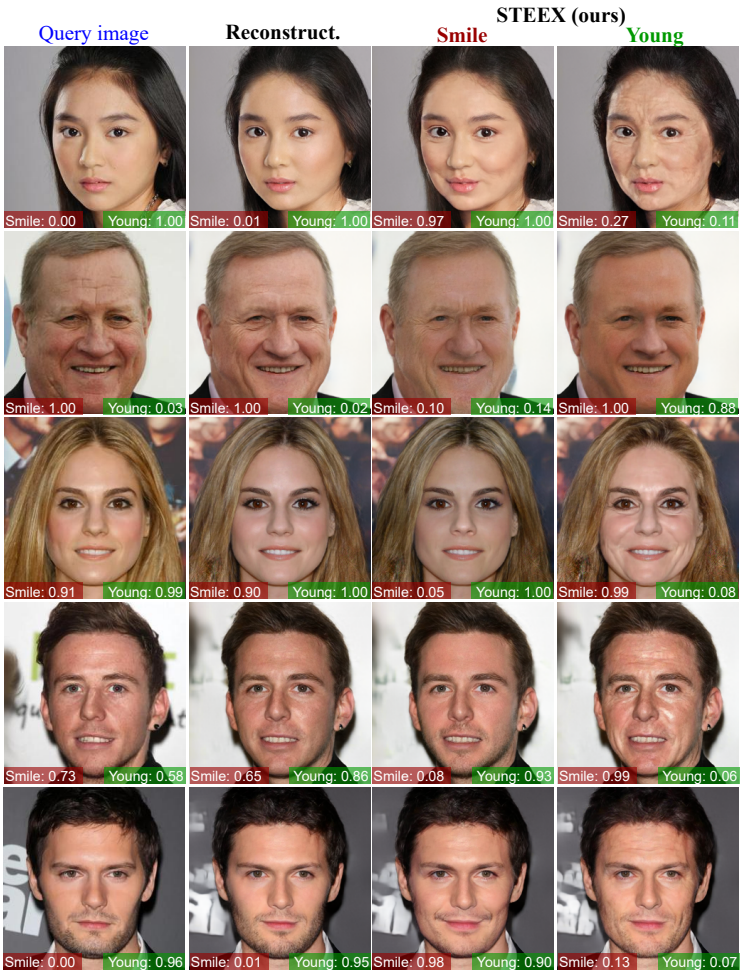

Fig. 2: Counterfactual explanations and reconstructions on CelebAMask-HQ generated by STEEX. Explanations are generated for two binary classifiers, on *Smile* and *Young* attributes, at resolution 256×256. Predicted scores are reported at the bottom of each image

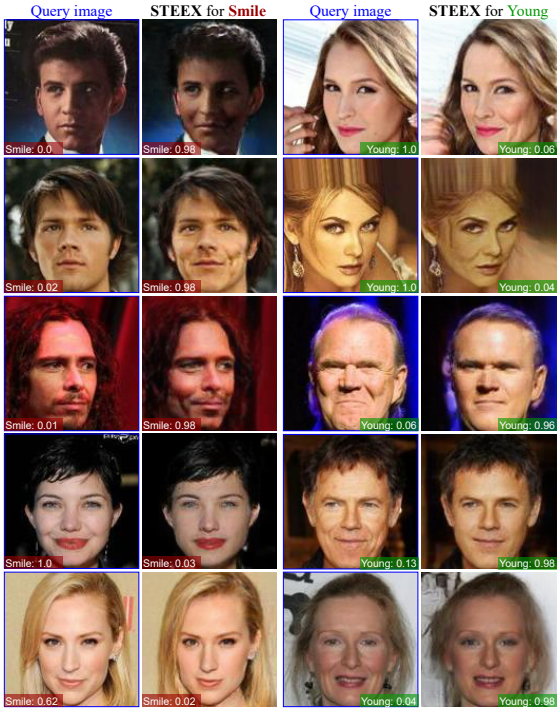

Fig. 3: Counterfactual explanations on CelebA generated by STEEX. Explanations are generated for two binary classifiers, on *Smile* and *Young* attributes, at resolution 128×128. Predicted scores are reported at the bottom of each image

model predicts '*Move forward*', several modifications can be observed depending on the image at hand, as reported by Fig. 5. 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. 6, to explain '*Move Forward*' decisions, by providing counterfactual images where the decision model predicts '*Stop*', modifications include green traffic 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.

STEEX vs. PE on BDD100k. In Fig. 7, we present a comparison between STEEX and PE [5] 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

Fig. 4: 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

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 best efforts, the adaptation of DiVE [4] 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.

STEEX on different decision models on CelebAMask-HQ. In Fig. 8, we show additional samples for the three different *Young*- classifiers on the CelebAMask-HQ dataset, with images of the size 256×256. Modifications found by STEEX hint at the specificities of each model: 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.

More details about the different models are given in Sec. C.

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*'

Fig. 5: 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

the image $G(S^I, z^I)$ generated from the predicted semantic mask $S^I = E_{\text{seg}}(x^I)$ 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. 1, we present a quantitative evaluation of the 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 and Tab. 2 of the main paper 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.

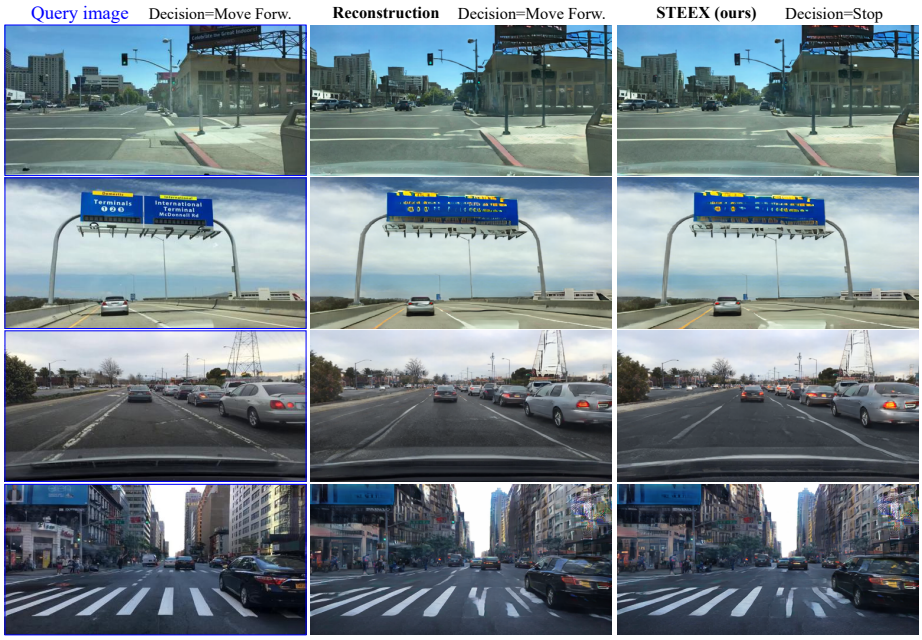

Fig. 6: Counterfactual explanations on BDD100k generated by STEEX, where the decision model initially predicts the '*Move forward*' class. Explanations are generated for a binary classifier trained with the BDD-OIA dataset extension annotated with the attribute *Move forward*. The image resolution is 512×256

In Fig. 1, Fig. 2, Fig. 5 and Fig. 6, we show some examples of reconstructions obtained by STEEX on CelebAMask-HQ and BDD100k. Overall, a reconstructed image is 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.

C Details on the analysis of decision models (Sec. 4.5)

The three different classifiers M_{top}, M_{mid}, and M_{bot}, presented in Sec. 4.5 are trained on modified images of the train set of CelebAMask-HQ where all pixels are masked out (with zeros) except for the top, middle, and bottom parts of the image respectively. More precisely, M_{top} only sees the top 65 pixel rows (out of 256), M_{bot} only keeps the bottom 56 pixel rows (out of 256) and M_{mid} only sees images where a centered rectangle of size 110×60. The decision model M_{full} is the one used for all other experiments, which is trained on unmodified images of the training set of CelebAMask-HQ. Note that the query image from the validation set on which the counterfactual explanation is provided is never modified. Model

Fig. 7: Counterfactual explanations on BDD100k generated by STEEX compared to explanations generated by Progressive Exaggeration (PE) [5]. All images have a 512×256 resolution

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

Table 1: 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

accuracies on the Young class are as follow: M_{full} : 89%, M_{top} : 83%, M_{mid} : 87%, M_{bot} : 86%.

D Technical Details

D.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.

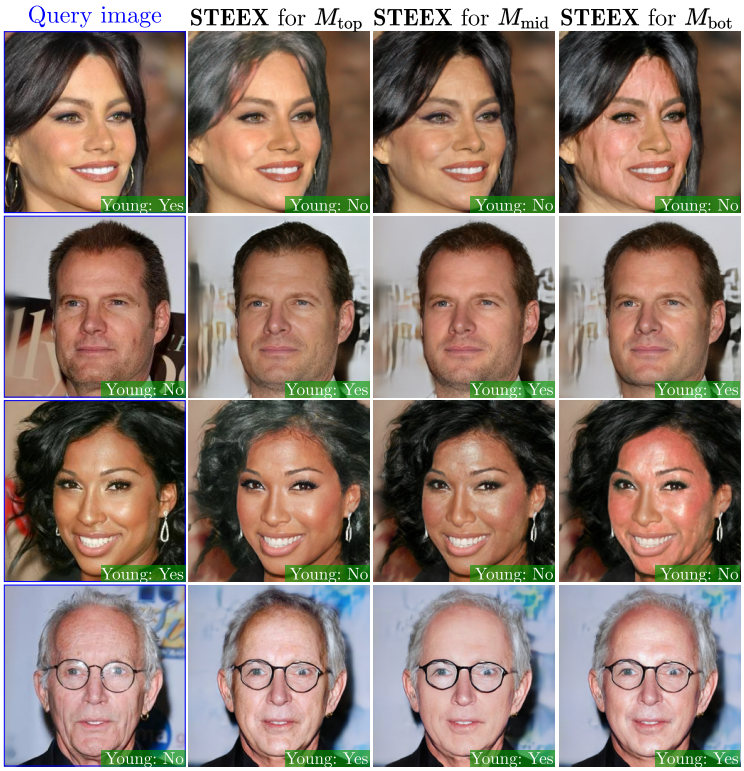

Fig. 8: Counterfactual explanations on CelebAMask-HQ for three different *Young* classifiers, namely M_{top}, M_{mid}, and M_{bot} that were trained on images where only the top, mid, and bottom parts of images were shown respectively. All images have a 256×256 resolution

D.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. 9, 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

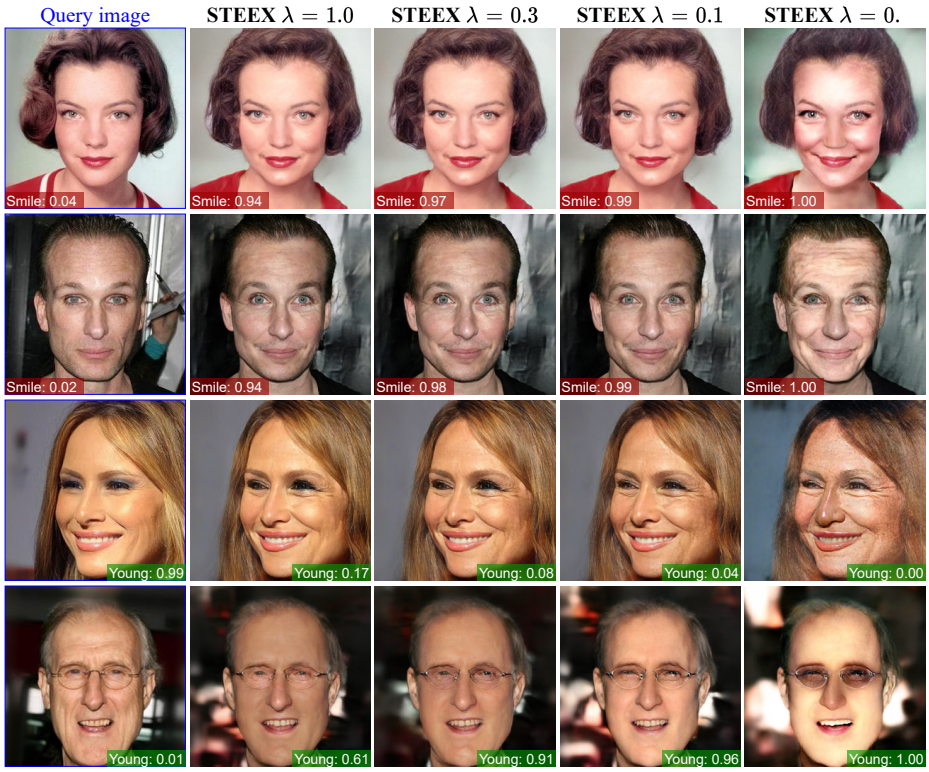

Fig. 9: Counterfactual explanations with various λ generated by STEEX. The λ parameter balances the contribution of the loss $\mathcal{L}_{\text{dist}}$ with respect to the one of $\mathcal{L}_{\text{decision}}$. When λ is high, the decision is 'lightly' changed and the counterfactual explanation remains 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

Progressive Exaggeration (PE) [5]. 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.

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
  $S^I \leftarrow E_{\text{seg}}(x^I)$ ▷ 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 $i \leftarrow 1$ to 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.
    $z \leftarrow \texttt{ADAM}(z, L, C, l_r)$ ▷ 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

D.3 Licenses

BDD100k data [7]. https://doc.bdd100k.com/license.html

BDD100k code. BSD 3-Clause License

BDD-OIA data [6]. No license provided

BDD-OIA code. BSD 3-Clause License

CelebA [3]. Agreement to use data on
https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

CelebAMask-HQ [2]. Agreement to use data on
https://mmlab.ie.cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html

SEAN [8] code. Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International https://github.com/ZPdesu/SEAN/blob/master/LICENSE.md

DiVE [4] code. Apache License 2.0
https://github.com/ElementAI/beyond-trivial-explanations/blob/master/LICENSE

PE [5] code. MIT Licence https://github.com/batmanlab/Explanation_by_Progressive_Exaggeration/blob/master/LICENSE.txt

DeepLabV3 [1] code. BSD 3-Clause License

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

References

1. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017) 12
2. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: CVPR (2020) 12
3. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015) 12
4. 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) 5, 12
5. Singla, S., Pollack, B., Chen, J., Batmanghelich, K.: Explanation by progressive exaggeration. In: ICLR (2020) 4, 8, 11, 12
6. 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) 11
7. 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) 11
8. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: SEAN: image synthesis with semantic region-adaptive normalization. In: CVPR (2020) 12