Supplementary Material
Text-to-Image Models for Counterfactual Explanations: a Black-Box Approach

Guillaume Jeanneret, Loïc Simon, Frédéric Jurie
Normandy University, ENSICAEN, UNICAEN, CNRS, GREYC, Caen, France
guillaume.jeanneret-sanmiguel@unicaen.fr

A. Evaluation Criteria

Before describing each metric and its formulation, we will thoroughly describe the goals of counterfactual explanations. As we stated in the main manuscript, counterfactual explanations seek to change an instance prediction by modifying the input instance. However, these modifications must be small but perceptually coherent. From the previous statement, we can extract many goals of CEs:

1. CEs must flip the decision of the classifier. In the literature, this feature is called *validity*.

2. The counterfactual changes should be plausible and realistic - simply referred to as *realism*. Visual automated systems are generally brittle to adversarial noise [1]. This noise is designed to fool the classifier, but with the restriction that it is hidden from visual inspection. Since this noise cannot be perceived, it cannot be analyzed to find spurious correlations. Therefore, only realistic and plausible changes are allowed.

3. The algorithm must generate *proximal and sparse* counterfactuals. One could create a valid and realistic explanation by simply replacing the target instance with a new one. This still obeys the realistic and valid goals. However, it does not give any information about the variables. Thus, the modifications must be sparse and close to the image to visually observe which variables have changed.

4. Finally, the algorithm must generate the explanation *efficiently*. This property is required to avoid delays for the user.

Now we will proceed to describe each evaluation metric and link it to its corresponding objective. As for notations, let $M(x, y)$ be the counterfactual algorithm applied to an image $x \in D$ targeting the class y, where D is a dataset. Additionally, let C be the classifier, $\mathbb{1}(condition)$ a function that is one if the condition is true or zero otherwise. Finally, let $a \in A$ be an attribute in a set A, then O^a is an attribute oracle classifier for a. This network predicts if its input has the attribute a. Similarly, let $\mathcal{O}$ be an identity verification network This DNN is trained to give a similarity measure between two images, often computed with the cosine similarity CS.

Success Rate. The success rate (or flip rate) measures the ratio at which counterfactuals have successfully reversed the original classifier's decision. This metric correlates with the validity goal. To measure it, we simply compute the proportion of valid counterfactuals to the size of the dataset, as in

$$SR = \frac{1}{|D|} \sum_{x \in D} \mathbb{1}(C(M(x, y)) = y). \tag{1}$$

Realism. To approximate the realism of the counterfactuals, the literature adopts the FID [2] metric from generation research. Furthermore, [5] extended the metric by computing the FID between the half of the dataset and the counterfactuals of the complement set. This was motivated to reduce the inherent bias in computing the FID, given that the difference between the original images and their CE is a few pixels in the image.

Proximity and Sparsity. To evaluate this goal, previous methods proposed several metrics to quantify the degree of dissimilarity between an instance and its explanation. Initially, most metrics were proposed for face images. Initially, [3] suggested using the mean number of attributes changed (MNAC), computed as follows:

$$MNAC = \frac{1}{|D|} \sum_{x \in D} \sum_{a \in A} \mathbb{1}(O^a(M(x, y)) \neq O^a(x)). \tag{2}$$

However, [4] noted that counterfactual methods will change some attributes if they are correlated. Thus, based on the MNAC, the Correlation Difference (CD) [4] measures the correlations produced by M. To further assess the proximity and sparsity in face counterfactuals, [8] suggested using the Face Verification Accuracy (FVA) to compute whether

M cannot modify the identity of the person. This metric is calculated as

$$FVA = \frac{1}{|D|} \sum_{x \in D} \mathbb{1}(CS(\mathcal{O}(x), \mathcal{O}(M(x, y))) > 0.5). \quad (3)$$

[5] noted that this metric was already saturated. To measure a more fine-grained metric, they proposed taking the continuous CS and calling the metric face similarity (FS):

$$FS = \frac{1}{|D|} \sum_{x \in D} CS(\mathcal{O}(x), \mathcal{O}(M(x, y)). \quad (4)$$

Finally, the same authors extended this metric for general-purpose images by computing Eq. 4 using a self-supervised trained model as $\mathcal{O}$. They called this metric S^3. Finally, [6] proposed to compute COUT. This metric computes the probability of the class y using multiple linear interpolations between x and $M(x, y)$.

Efficiency. The literature generally ignores computing an *efficiency* metric. To compute the efficiency of counterfactual models, the widely accepted metric is floating point operations (FLOPs). In addition, it is also recommended to compute the average time per counterfactual. However, this metric is only comparable if all measurements are computed on the under the same circumstances.

B. Qualitative Results

In this section, we provide additional qualitative results. For the CelebA HQ [7] dataset, we provide our and ACE [5] counterfactuals to show the differences.

References

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[2] 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. *Advances in neural information processing systems*, 30, 2017. 1

[3] Paul Jacob, Éloi Zablocki, Hédi Ben-Younes, Mickaël Chen, Patrick Pérez, and Matthieu Cord. Steex: steering counterfactual explanations with semantics. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 387–403. Springer, 2022. 1

[4] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision*, pages 858–876, 2022. 1

[5] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16425–16435, 2023. 1, 2

[6] Saeed Khorram and Li Fuxin. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10203–10212, 2022. 2

[7] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[8] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. In *International Conference on Learning Representations*, 2020. 1

Figure 1. Counterfactual Explanations targeting the Non-Smile attribute.

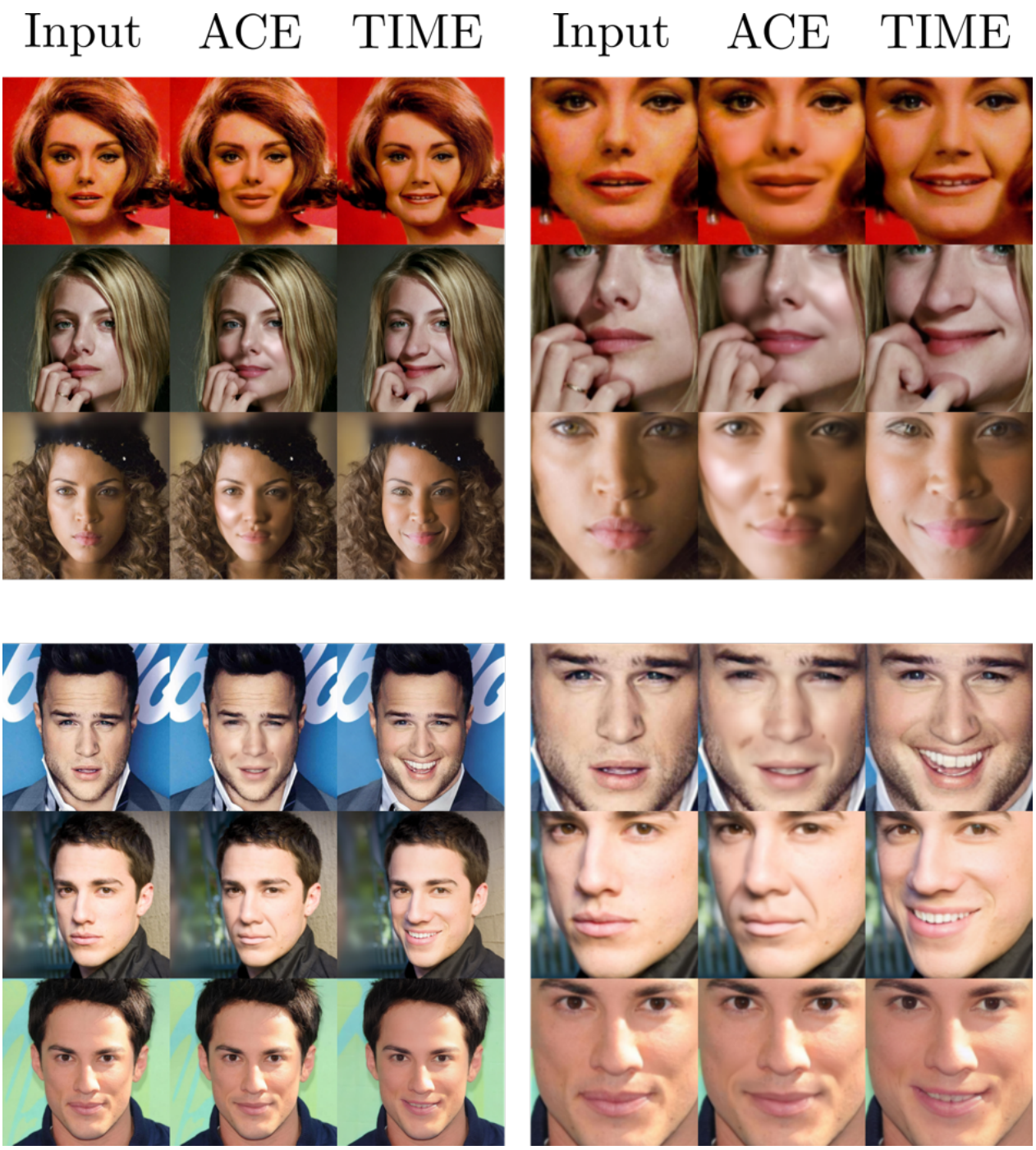

Figure 2. Counterfactual Explanations targeting the Smile attribute.

Figure 3. Counterfactual Explanations targeting the Young attribute.

Figure 4. Counterfactual Explanations targeting the Old attribute.

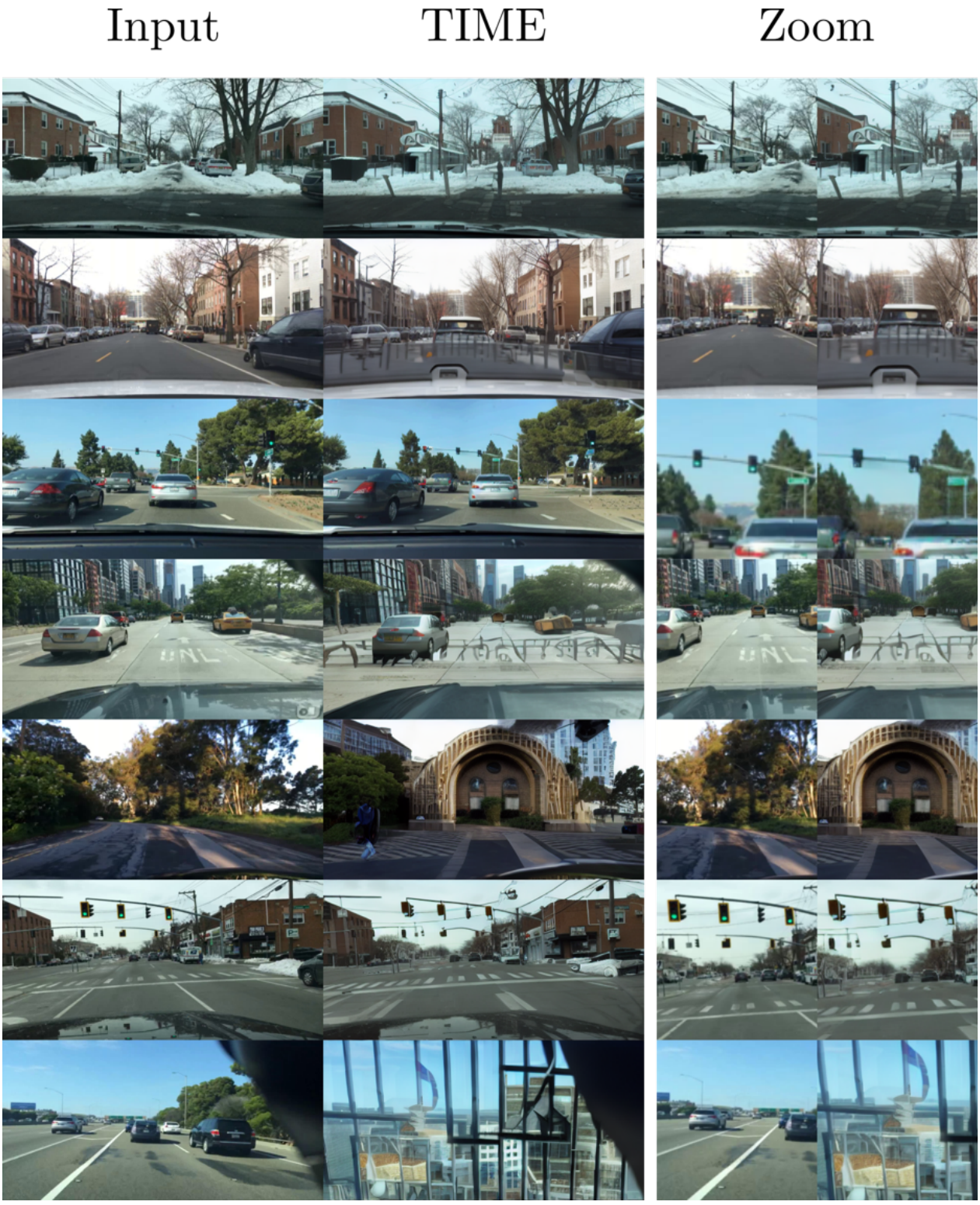

Figure 5. Counterfactual Explanations targeting the Stop action.

Figure 6. Counterfactual Explanations targeting the Forward action.