Adversarial Counterfactual Visual Explanations

Guillaume Jeanneret, Loïc Simon, Frédéric Jurie

Adversarial Counterfactual Visual Explanations

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

Abstract

Counterfactual explanations and adversarial attacks have a related goal: flipping output labels with minimal perturbations regardless of their characteristics. Yet, adversarial attacks cannot be used as is in a counterfactual explanation perspective, as such perturbations are perceived as noise and not as actionable and understandable image modifications. Building on the robust learning literature, this paper proposes an elegant method to turn adversarial attacks into semantically meaningful perturbations, without modifying the classifiers to explain. The proposed approach hypothesizes that Denoising Diffusion Probabilistic Models are excellent regularizers for avoiding high-frequency and out-of-distribution perturbations when generating adversarial attacks. The paper's key idea is to build attacks through a diffusion model to polish them, which allows studying a model regardless of its robustification level. Extensive experimentation shows the advantages of our counterfactual explanation approach over current State-of-the-Art in multiple testbeds.

1. Introduction

The research branch of explainable artificial intelligence has yielded remarkable results, gradually opening the machine learning black boxes. The production of counterfactual explanations (CE) has become one of the promising pipelines for explainability, especially in computer vision [25, 28, 49, 54]. As a matter of fact, CE are an intuitive way to expose how an input instance can be minimally modified to steer the desired change in the model's output. More precisely, CE answers the following: *what does X have to change to alter the prediction from Y to Y'?* From a user perspective, these explanations are easy to understand since they are concise and illustrated by examples. Henceforth, companies have adopted CE as an interpretation methodology to legally justify the decision-making of machine learning models [60]. To better appreciate the potential of CE, one may consider the following scenario: a client goes to a photo booth to take some ID photos, and the system claims the photos are invalid for such usage. Instead of performing random attempts to abide by the administration criteria, an approach based on CE could provide visual indications of what the client should fix.

The main objective of CE is to add minimalistic semantic changes in the image to flip the original model's prediction. Yet, these generated explanations must accomplish several objectives [28,49,60]. A CE must be *valid*, meaning that the CE has to change the prediction of the model. Secondly, the modifications have to be *sparse and proximal* to the input data, targeting to provide simple and concise explanations. In addition, the CE method should be able to generate *diverse* explanations. If a trait is the most important for a certain class among other features, diverse explanations should change this attribute most frequently. Finally, the semantic changes must be *realistic*. When the CE method inserts out-of-distribution artifacts in the input image, it is difficult to interpret whether the flipping decision was because of the inserted object or because of the shifting of the distribution, making the explanation unclear.

Adversarial attacks share a common goal with CE: flipping the classifier's prediction. For traditional and non-robust visual classifiers, generating these attacks on input instances creates imperceptible noise. Even though it has been shown that it contains meaningful changes [24] and that adversarial noise and counterfactual perturbations are related [13, 23], adversarial attacks have lesser value. Indeed, the modifications present in the adversaries are unnoticeable by the user and leave him with no real feedback.

Contrary to the previous observations, many papers (*e.g.*, [47]) evidenced that adversarial attacks toward *robust* classifiers generate semantic changes in the input images. This has led works [51, 70] to explore robust models to produce data using adversarial attacks. In the context of counterfactual explanations, this is advantageous [5, 52] because the optimization will produce semantic changes to induce the flipping of the label.

Then two challenges arise when employing adversarial attacks for counterfactual explanations. On the one hand, when studying a classifier, we must be able to explain its behavior regardless of its characteristics. So, a naive ap-

plication of adversarial attacks is impractical for non-robust models. On the other hand, according to [57], robustifying the classifier yields an implicit trade-off by lowering the *clean accuracy*, as referred by the adversarial robustness community [10], a particularly crucial trait for high-stakes areas such as the medical field [40].

The previous remarks motivate our endeavor to mix the best of both worlds. Hence, in this paper, we propose robustifying brittle classifiers *without* modifying their weights to generate CE. This robustification, obtained through a simple filtering preprocessing leveraging diffusion models [19], allows us to keep the performance of the classifier untouched and unlocks the production of CE through adversarial attacks.

We summarize the novelty of our paper as follows: (i) We propose Adversarial Counterfactual Explanations, ACE in short, a novel methodology based on adversarial attacks to generate semantically coherent counterfactual explanations. (ii) ACE performs competitively with respect to the other methods, beating previous state-of-the-art methods in multiple measurements along multiple datasets. (iii) Finally, we point out some defects of current evaluation metrics and propose ways to remedy their shortcomings. (iv) To show a use case of ACE, we study ACE's meaningful and plausible explanations to comprehend the mechanisms of classifiers. We experiment with ACE findings producing actionable modifications in real-world scenarios to flip the classifier decision.

To promote the research in counterfactual explanations, we will make our code available upon acceptance.

2. Related Work

Explainable AI. The main dividing line between the different branches of explainable artificial intelligence stands between *Ad-Hoc* and *Post-Hoc* methods. The former promotes architectures that are interpretable by design [3, 4, 21, 50] while the latter considers analyzing existing models as they are. Since our setup lies among the Post-Hoc explainability methods, we spotlight that this branch splits into global and local explanations. The former explains the general behavior of the classifier, as opposed to a single instance for the latter. This work belongs to the latter. There are multiple local explanations methods, from which we highlight saliency maps [8, 26, 32, 35, 61, 68], concept attribution [15, 31, 33] and model distillation [14, 55]. Concisely, these explanations try to shed light on *how* a model took a specific decision. In contrast, we focus on the ongrowing branch of counterfactual explanations, which tackles the question: *what* does the model uses for a forecast? We point out that some novel methods [17, 59, 62, 63] call themselves counterfactual approaches. Yet, these systems highlight regions between a pair of images without producing any modification.

Counterfactual Explanations. CE have taken momentum in recent years to explain model decisions. Some methods rely on prototypes [37] or deep inversion [56], while other works explore the benefits of other classification models for CE, such as Invertible CNNs [22] and Robust Networks [5, 52]. A common practice is using generative tools as they give multiple benefits when producing CE. In fact, using generation techniques is helpful to generate data in the image manifold. There are two modalities to produce CE using generative approaches. On the one hand, many methods use conditional generation techniques [37, 54, 58] to fit what a classification model learns or how to control the perturbations. On the other hand, unconditional approaches [28, 30, 44, 49, 53, 67] optimize the latent space vectors. Among the counterfactual approaches, we draw attention to Jeanneret *et al*. [28]'s work. This method uses a modified version of the guided diffusion [12] to steer the generation toward the target label. In contrast, even when we use DDPM, we use adversarial attacks directly on the image space to generate semantic changes before post-processing it through the diffusion model without relying on controlling the generation process.

Adversarial Attacks and their relationship with CE. Adversarial attacks share the same main objective as counterfactual explanations: flipping the forecast of a target architecture. On the one hand, *white-box* attacks [7, 10, 16, 27, 39, 42] leverage the gradients of the input image with respect to a loss function to construct the adversary. Also, universal noises [41] are adversarial perturbations created for fooling many different instances. On the other hand, *black-box* attacks [2, 48, 69] restrain their attack by checking merely the output of the model. Finally, Nie *et al*. [45] study DDPMs from a robustness perspective, disregarding the benefits of counterfactual explanations.

In the context of CE for visual models, the produced noises are indistinguishable for humans when the network does not have any defense mechanism, making them useless. This lead works [1, 23, 46] to approach the relationship between these two research fields. Compared to previous approaches, we manage to leverage adversarial attacks to create semantic changes in undefended models to explore their semantic weaknesses perceptually in the images; a difficult task due to the nature of the data.

3. Adversarial Counterfactual Explanations

The key contribution of this paper is the Adversarial Counterfactual Explanations (ACE) method. ACE produces counterfactual images in two steps, as seen in Figure 1. We briefly introduce these two steps here and detail them in the following sections.

Step 1. Producing pre-explanation images (§3.1). Let $L_{class}(x; y)$ be a function measuring the agreement between the sample x and class y. This function is typically

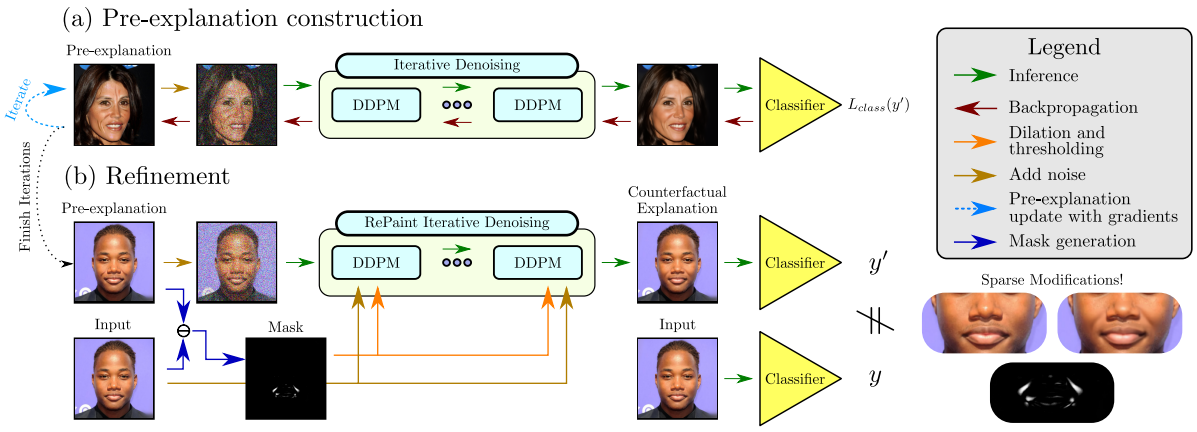

Figure 1. **Pre-explanation Construction and Refinement** ACE generates the counterfactual explanation in a two-step sequence. Initially, (a) To generate semantic updates in the input image, the DDPM processes the instance before computing the loss function $L_{class}(y')$, where y' is the target label. Then, it uses the gradients of the image with respect to the criterion to update it. After all the iterative updates (b) we generate a binary mask using the magnitude's difference between the explanation and input image. Leveraging this mask, we refine the pre-explanation using RePaint's inpainting method.

the cross-entropy loss of the classifier we are studying with respect to y. With ACE, generating the pre-explanation image of (x, y) for the target class $y' \neq y$ consists in finding x' minimizing $L_{class}(F(x'); y')$. Here, $F(x')$ is a filtering function that constrains the attack to stay in the manifold of the training images. In a nutshell, the filtering process F robustifies the fragile classifier under examination to generate semantic changes *without* modifying its weights.

Step 2. Bringing the pre-explications closer to the input images (§3.2). The pre-explanation generation restricts only those pixels in the image that are useful in switching the output label from y to y'. The rest of the pixels are only implicitly constrained by the design of F. Accordingly, the purpose of this second step is to keep these non-explicitly constrained pixels identical to those of the input image.

3.1. Pre-explanation generation with DDPMs

To avoid generating adversarial noise and producing useful semantics, the previously introduced function F should have two key properties. (i) Removing high-frequency information that traditional adversarial attacks generate. Indeed, these perturbations could change the classifier's decision without being actionable or understandable by a human. (ii) Producing in-distribution images without distorting the input image. This property seeks to maintain the image structures not involved in the decision-making process as similar as possible while avoiding giving misleading information to the user.

Denoising Diffusion Probabilistic Models [19], commonly referred to as DDPM or diffusion models, achieve these properties if used properly. On the one hand, each inference through the DDPM is a denoising process; in particular, it removes high-frequency signals. On the other hand, DDPMs generate in-distribution images.

As a reminder, DDPMs rely on two Markov chains, one inverse to the other. The forward chain *adds* noise from a state t into $t+1$ while the reverse chain *removes* Gaussian noise from $t+1$ to t. Noting x_t the instance at time step t, the forward chain is directly simulated from a clean instance x_0 through

$$x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1-\alpha_t}\, \epsilon,\ \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where α_t is a time-dependent constant. At inference, the DDPM produces a mean $\mu_t(x_t)$ and a deviation matrix $\Sigma_t(x_t)$. Using these variables, the next less noisy image is sampled from

$$x_{t-1} = \mu_t(x_t) + \Sigma_t(x_t)\, \epsilon,\ \epsilon \sim \mathcal{N}(0, I). \tag{2}$$

Thus, the DDPM denoising algorithm iterates the previous step until $t = 0$ arriving at an image without noise. Please refer to previous works [12, 19] for a thorough understanding of diffusion models.

ACE pre-explanation generation. Starting from a query image x, we can obtain a filtered version by applying the forward DDPM process up to level τ (Eq. 1) and then denoise it iteratively thanks to the iterative DDPM denoising steps from level $t = \tau$ (Eq 2). In this case, to highlight the use of this intermediate step τ, we denote the diffusion filtering process as $F = F_\tau$ (Figure 1a). Thus, we optimize the image through the DDPM filtering process, F_τ, before computing the classification loss. So, we obtain the pre-explanations with

$$\underset{x'}{\operatorname{argmin}}\ L_{class}(F_\tau(x'); y') + \lambda_d\, d(x', x), \tag{3}$$

where λ_d is a regularization constant and d a distance function.

3.2. Bringing the pre-explanations closer to the input images

By limiting the value of τ, the DDPM will not go far enough to generate a normal distribution, and the reconstruction will somehow preserve the overall structure of the image. However, we noted that a post-processing phase could help keep irrelevant parts of the image untouched. For example, in the case of faces, the denoising process may change the hairstyle while targeting the smile attribute. Since hairstyle is presumably uncorrelated with the smile feature, the post-process should neutralize those unnecessary alterations.

To this end, we first compute a binary mask m delineating regions that qualify for modifications. To do so, we consider the magnitude difference between the pre-explanation and the original mask, we dilate this gray-scale image and threshold it, yielding the desired mask. This matter being settled, we need to fuse the CE inside the mask along with the input outside the mask.

In that aim, a natural strategy is using inpainting methods. So, we leverage RePaint's recent technique [38], originally designed for image completion, and adapt it to our *picture-in-picture* problem (Figure 1b). This adaptation is pretty simple and integrates very well with the rest of our framework. It starts from the noisy pre-explanations x_τ and iterate the following altered denoising steps:

$$x_{t-1} = \mu_t(x'_t) + \Sigma_t(x'_t)\,\epsilon,\ \epsilon \sim \mathcal{N}(0, I). \tag{4}$$

where $x'_t = x_t \cdot m + x^i_t \cdot (1 - m)$ is the raw collage of the current noisy reconstruction x_t and the noisy version x^i_t of the initial instance at the same noise level t obtained with Eq. 1. At the end of this process, the final image x_0 will be identical to the input sample outside of the mask, and very similar to the CE within the mask.

4. Experimentation

4.1. Evaluation Protocols and Datasets

Datasets. In line with the recent literature on counterfactual images [28, 29, 49, 54], first, we evaluate ACE on CelebA [36], with images of size of 128×128 and a DenseNet121 classifier [20], for the 'smile' and 'age' attributes. Following Jacob *et al*. [25], we experimented on CelebA HQ [34] and BDD100k [65]. CelebA HQ has a higher image resolution of 256×256. BDD100k contains complex traffic scenes as 512×256 images; the targeted attribute is 'forward' vs 'slow down'. The decision model is also a DenseNet121, trained on the BDD-IOA [64] extension dataset. Regarding the classifiers for which we want to generate counterfactuals, we took the pre-trained weights from DiME [28] open source for CelebA and from STEEX [25] for CelebA HQ and BDD100k, for fair comparisons.

Evaluation criteria for quantitative evaluation.
Validity of the explanations is commonly measured with the Flip Rate (FR), *i.e*. how often the CE is classified as the target label.
Diversity is measured by extending the diversity assessment from Mothilal *et al*. [43]. As suggested by Jeanneret *et al*. [28], the diversity is measured as the average LPIPS [66] distance between pairs of counterfactuals.
Sparsity or proximity has been previously evaluated with several different metrics [49, 54], in the case of face images and face attributes. On the one hand, the mean number of attributes changed (MNAC) measures the smallest amount of traits changed between the input-explanation pair. Similarly, this metric leverages an oracle network pretrained on VGGFace2 [6] and then fine-tuned on the dataset. Further, Jeanneret *et al*. [28] showed the limitations of the MNAC evaluation and proposed the CD metric to account for the MNAC's limitations. On the other hand, to measure whether an explanation changed the identity of the input, the assessment protocol uses face verification accuracy [6] (FVA). To this end, the evaluation uses a face verification network. However, FVA has 2 main limitations: i) it can be applied to face related problems only, ii) it works at the level of classifier decisions which turns out to be too rough when comparing an image to its CE, as it involves only a minimal perturbation. For face problems, we suggest skipping the thresholding and consider the mean cosine distance between the encoding of image-counterfactual pairs, what we refer to as Face Similarity (FS). To tackle non-face images, we propose to extend FS by relying on self-supervised learning to encode image pairs. To this end, we adopted SimSiam [9] as an encoding network to measure the cosine similarity. We refer to this extension as SimSiam Similarity (S^3). Finally, also for classifiers that are not related to faces, Khorram *et al*. [30] proposed COUT to measure the transition probabilities between the input and the counterfactual.
Realism of counterfactual images [54] is usually evaluated by the research community with the FID [18] between the original set and the valid associated counterfactuals. We believe there is a strong bias as most of the pixels of counterfactuals are untouched and will dominate the measurement, as observed in our ablation studies (Sec. 4.6). To remove this bias, we split the dataset into two sets, generating the CE for one set and measuring the FID between the generated explanations and the other set, iterating this process ten times and taking the mean. We call this metric sFID.

Implementation details. One of the main obstacles of diffusion models is transferring the gradients through all the iterations of the iterative denoising process. Fortunately, diffusion models enjoy a time-step re-spacing mechanism, allowing us to reduce the number of steps at the cost of a quality reduction. So, we drastically decreased the number of sampling steps to construct the pre-explanation. For

	Smile							Age						
Method	FID	sFID	FVA	FS	MNAC	CD	COUT	FID	sFID	FVA	FS	MNAC	CD	COUT
DiVE	29.4	-	97.3	-	-	-	-	33.8	-	*98.2*	-	4.58	-	-
DiVE^{100}	36.8	-	73.4	-	4.63	2.34	-	39.9	-	52.2	-	4.27	-	-
STEEX	10.2	-	96.9	-	4.11	-	-	11.8	-	97.5	-	3.44	-	-
DiME	3.17	4.89	98.3	0.729	3.72	2.30	0.5259	4.15	5.89	95.3	0.6714	*3.13*	3.27	0.4442
ACE ℓ_1	**1.27**	**3.97**	**99.9**	**0.874**	*2.94*	*1.73*	**0.7828**	**1.45**	**4.12**	**99.6**	*0.7817*	3.20	*2.94*	**0.7176**
ACE ℓ_2	*1.90*	*4.56*	**99.9**	*0.867*	**2.77**	**1.56**	*0.6235*	*2.08*	*4.62*	**99.6**	**0.7971**	**2.94**	**2.82**	*0.5641*

Table 1. **CelebA Assessment.** Main results for CelebA dataset. We extracted the results from DiME and STEEX papers. In **bold** and *italic* we show the best and second-best performances. ACE outperforms all methods in every assessment protocol.

CelebA [36], we instantiate the DDPM [12] model using DiME's [28] weights. In practice, we set $\tau = 5$ out of 50 steps. For CelebA HQ [34], we fixed the same τ, but we used the re-spaced time steps to 25 steps. For BDD100k [65], we follow the same settings as STEEX [25]: we trained our diffusion model on the 10.000 image subset of BDD100k. To generate the explanations, we used 4 steps out of 100. Additionally, all our methods achieve a success ratio of 95% at minimum. We will detail in the supplementary material all instructions for each model on every dataset. We adopted an ℓ_1 or ℓ_2 distance for the distance function. Finally, for the attack optimization, we chose the PGD [39] without any bound and with 50 optimization steps.

4.2. Comparison Against the State-of-the-Art

In this section, we quantitatively compare ACE against previous State-of-the-Art methods. To this end, we show the results for CelebA [36] and CelebA HQ [34] datasets in Table 1 and Table 2, respectively. Additionally, experiments on the BDD100k [65] dataset are given in Table 3. To extend the study of BDD, we further evaluated our proposed approach on the BDD-IOA [64] validation set, also presented in Table 3. Since DiME has shown superior performance over the previous literature [29, 49, 54], we compare only to DiME.

DiME experimented originally on CelebA only. Hence, they did not tune their parameters for CelebA HQ and BBD100k. By running their default parameters, DiME achieves a flip rate of 41% in CelebA HQ. We fix this by augmenting the scale hyperparameter for their loss function. DiME's new success rate is 97% for CelebA HQ. For BDD100k, our results showed that using fewer steps improves the quality. Hence, we used 45 steps out of their re-spaced 200 steps. Unfortunately, we only managed to increase their success ratio to 90.5%.

These experiments show that the proposed methodology beats the previous literature on most metrics for all datasets. For instance, ACE, whatever the chosen distance, outmatches DiME on all metrics in CelebA. For the CelebAHQ, we noticed that DiME outperforms ACE only for the COUT and CD metrics. Yet, our proposed method remains comparable to theirs. For BDD100k, we remark that our method consistently outperforms DiME and STEEX.

Two additional phenomena stand out within these results. On the one hand, we observed that the benefit of favoring ℓ_1 over ℓ_2 depends on the characteristics of the target attribute. We noticed that the former generates sparser modifications, while the latter tends to generate broader editing. This makes us emphasize that different attributes require distinct modifications. On the other hand, these results validate the extensions for the FVA and FID metrics. Indeed, the difference between the FVA values on CelebA are small (from 98.3 to 99.9). Yet, the FS shows a major increase. Additionally, for the Age attribute on CelebA HQ, ACE ℓ_2 shows a better performance than DiME for the FID metric. The situation is reversed with sFID as DiME is slightly superior.

To complement our extensive experimentation, we tested ACE on a small subset of classes on ImageNet [11] with a ResNet50. We selected three pairs of categories for the assessment, and the task is to generate the CE targeting the contrary class. For the FID computation, we used only the instances from both categories but not external data since we are evaluating the in-class distribution.

We show the results in Table 4. Unlike the previous benchmarks, ImageNet is extremely complex and the classifier needs multiple factors for the decision-making process. Our results reflect this aspect. We believe that current advancements in CE still need an appropriate testbed to validate the methods in complex datasets such as ImageNet. For instance, the model uses the image's context for forecasting. So, choosing the target class without any previous information is unsound.

4.3. Diversity Assessment

In this section, we explore ACE's ability to generate diverse explanations. Diffusion models are, by design, capable of generating distributions of images. Like [28], we take advantage of the stochastic mechanism to generate perceptually different explanations by merely changing the noise for each CE version. Additionally, for a fair comparison,

	Smile							Age						
Method	FID	sFID	FVA	FS	MNAC	CD	COUT	FID	sFID	FVA	FS	MNAC	CD	COUT
DiVE	107.0	-	35.7	-	7.41	-	-	107.5	-	32.3	-	6.76	-	-
STEEX	21.9	-	*97.6*	-	5.27	-	-	26.8	-	96.0	-	5.63	-	-
DiME	18.1	27.7	96.7	0.6729	2.63	**1.82**	**0.6495**	18.7	*27.8*	*95.0*	0.6597	2.10	*4.29*	**0.5615**
ACE ℓ_1	**3.21**	**20.2**	**100.0**	**0.8941**	**1.56**	2.61	0.5496	**5.31**	**21.7**	**99.6**	**0.8085**	**1.53**	5.4	0.3984
ACE ℓ_2	*6.93*	*22.0*	**100.0**	*0.8440*	*1.87*	*2.21*	*0.5946*	*16.4*	28.2	**99.6**	*0.7743*	*1.92*	**4.21**	*0.5303*

Table 2. **CelebAHQ Assessment.** Main results for CelebA HQ dataset. We extracted the results from STEEX's paper. In **bold** and *italic* we show the best and second-best performances, respectively. ACE outperforms most methods in many assessment protocols.

Method	FID	sFID	S^3	COUT	FR
BDD-OIA					
DiME	13.70	26.06	0.9340	0.3188	91.68
ACE ℓ_1	**2.09**	**22.13**	**0.9980**	*0.7404*	99.91
ACE ℓ_2	*3.3*	*22.75*	*0.9949*	**0.7840**	100.0
BDD100k					
STEEX	58.8	-	-	-	99.5
DiME	7.94	11.40	0.9463	0.2435	90.5
ACE ℓ_1	**1.02**	**6.25**	**0.9970**	*0.7451*	99.9
ACE ℓ_2	*1.56*	*6.53*	*0.9946*	**0.7875**	99.9

Table 3. **BDD Assessement.** Main results for BDDOIA and BDD100k datasets. We extracted STEEX's results from their paper. In **bold** and *italic* we show the best and second-best performances, respectively.

Method	FID	sFID	S^3	COUT	FR
Zebra – Sorrel					
ACE ℓ_1	84.5	122.7	0.9151	-0.4462	47.0
ACE ℓ_2	67.7	98.4	0.9037	-0.2525	81.0
Cheetah – Cougar					
ACE ℓ_1	70.2	100.5	0.9085	0.0173	77.0
ACE ℓ_2	74.1	102.5	0.8785	0.1203	95.0
Egyptian Cat – Persian Cat					
ACE ℓ_1	93.6	156.7	0.8467	0.2491	85.0
ACE ℓ_2	107.3	160.4	0.7810	0.3430	97.0

Table 4. **ImageNet Assessement.** We test our model in ImageNet. We generated the explanations for three sets of classes. Producing CE for these classes remains a challenge.

we do not use the RePaint's strategy here because DiME does not have any local constraints and can, as well, change useless structures, like the background. To validate our approach, we follow [28] assessment protocol. Numerically, we obtain a diversity score of 0.110 while DiME reports 0.213. Since DiME corrupts the image much more than ACE even without RePaint, the diffusion model has more opportunities to generate distinct instances. In contrast, we do not go deep into the forward noising chain to avoid changing the original class when performing the filtering.

To circumvent the relative lack of diversity, we vary the re-spacing at the refinement stage and the sampled noise. Note that later in the text, we show that using all steps without any re-spacing harms the success ratio. So, we set the new re-spacing such that it respects the accuracy of counterfactuals and fixed the variable number of noise to maintain the ratio between τ and the re-spaced number of sampling steps ($5/50$ in this case). Our diversity score is then of 0.1436. Nevertheless, DiME is better than ACE in terms of diversity, but this is at the expense of the other criteria, because its diversity comes, in part, from regions of the images that should not be modified (for example, the background).

4.4. Qualitative Results

We show some qualitative results in Figure 2 for all datasets and included some ImageNet examples. From an attribute perspective, some have sparser or coarser characteristics. For instance, age characteristics cover a wider section of the face, while the smile attribute is mostly located in small regions of the image. Our qualitative results expose that different distance losses impose different types of explanations. For this case, ℓ_1 loss exposes the most local and concrete explanations. On the other hand, the ℓ_2 loss generates coarser editing. This feature is desired for certain classes, but it is user-defined. Additionally, we note that the generated mask is useful to spot out the location of the changes. This is advantageous as it exemplifies which changes were needed and where they were added. Most methods do not indicate the localization of the changes, making them hard to understand. In the supplementary material, we will include more qualitative results.

4.5. Actionability

Counterfactual explanations are expected to teach the user plausible modifications to change the classifier's prediction. In this section, we study a batch of counterfactual-input tuples generated with our method. If ACE is capable of creating useful counterfactual explanations, we should be qualified to understand some weaknesses or some behaviors of our classifier. Additionally, we should be able to fool the classifier by creating the necessary changes in real life. To this end, we studied the CelebA HQ classifier for the age

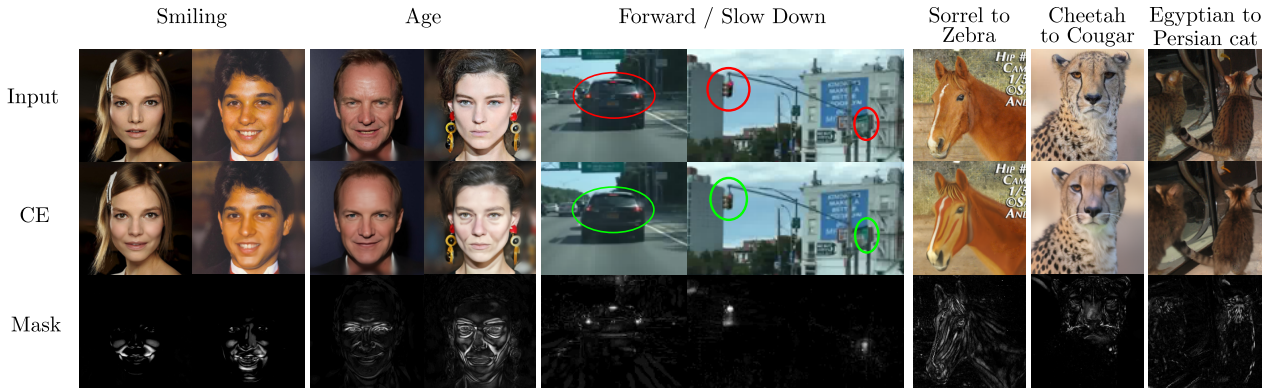

Figure 2. **Qualitative Results.** ACE create sparse but realistic changes in the input image. Further, ACE enjoys from the generate mask, which helps in understanding which and where semantic editing were added. The first row displays the input images, the second one the counterfactual explanations and the third the corresponding mask.

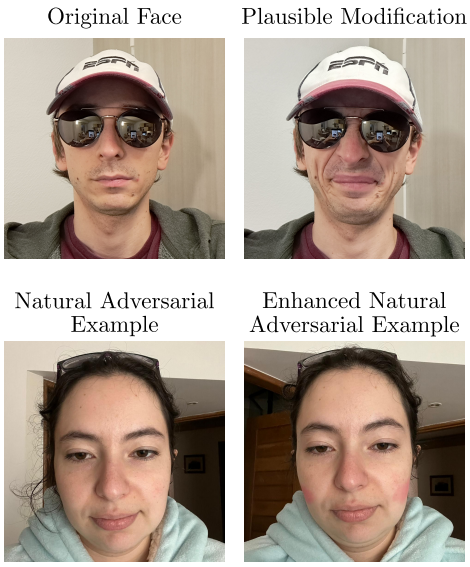

Figure 3. **Actionability.** From browsing our counterfactuals, we found two weaknesses of the scrutinized classifier. Row 1: We tested if a frown could change the classification from young to old. Row 2: we checked if having high cheekbones flipped is enough to classify someone as smiling. Both experiments were successful.

and smile attributes.

After surveying some images and their explanations, we identified two interesting results (Figure 3). Many of the counterfactual explanations changing from 'young' to 'old' evidence that frowning could change the prediction of the classifier. So, we tested this hypothesis in the real life. We took a photo one individual before and after the frown, avoiding changing the scenery. We were successful and managed to change the prediction of the classifier. For smile, we identified a spurious correlation. Our counterfactuals show that the classifier uses the morphological trait of high cheekbones to classify someone as smiling as well as having red cheeks. So, we tested whether the classification model wrongly predicts as smiling someone with high cheekbones even when this person is not smiling. We also tested whether we can enhance it with some red make up in the cheeks. Effectively, our results show that having high cheekbones is a realistic adversarial feature toward the smiling attribute for the classifier. Also, the classifier confidence (probability) can be strengthened by adding some red make up in the cheeks. These examples demonstrate the applicability of ACE in real scenarios.

4.6. Ablation Studies

In this section, we scrutinize the differences between the pre-explanation and the refined explanations. Then we explore the effects of using other types of adversarial attacks. Finally, we show that the S^3 metric gives similar results as the FVA, as a sanity check.

Pre-Explanation *vs* Counterfactual Explanations. We explore here, quantitatively and qualitatively, the effects of the pre-explanations (Pre-CE). Additionally, we use the diffusion model without any inpainting strategy to filter the explanations with (FR-CE) and without (F-CE) the re-spacing method. Finally, we compare them against the complete model (ACE). To quantitatively compare all versions, we conducted this ablation study on the CelebA dataset for both 'smile' and 'age' attributes. We assessed the components using the FID, sFID, MNAC, CD, and FR metrics. Note that, we did not include the FVA or FS metrics in this assessment, as these values did not vary much and do not provide insightful information; the FVA is $\sim$99.9 and FS $\sim$0.87 for all versions.

We show the results in Table 5. We observe that pre-explanations have a low FID. Nonetheless, their sFID is worse than the F-CE version. As said before, we noticed that including both input and counterfactual in the FID assessment introduces a bias in the final measurement, and

Method	FID	sFID	MNAC	CD	FR
Smile					
Pre-CE	1.87	4.63	3.48	3.05	99.82
FR-CE	8.31	10.30	3.43	1.68	99.97
F-CE	2.64	4.61	3.16	1.56	93.37
ACE	1.27	3.97	2.94	1.73	99.86
Age					
Pre-CE	3.93	6.71	3.76	3.17	99.55
FR-CE	7.10	9.09	3.13	2.66	99.77
F-CE	4.23	6.20	3.53	3.04	93.50
ACE	2.08	4.62	2.94	2.82	99.35

Table 5. **Refinement Ablation.** We show the importance of each component from ACE. FR stands for flip rate.

this experiment confirms this phenomenon. Additionally, one can check that the MNAC metric between the pre-explanation and the FR-CE version does not vary much, yet, the CD metric for the FR-CE is much better. This evidences that the generative model can capture the dependencies between the attributes. Also, we notice that the flip rate (FR) is much lower when using all diffusion steps instead of the re-spaced alternative. We expected this behavior, since we create the pre-explanation to change the classifier's prediction with re-spaced time steps within the DDPM.

Qualitatively, we point out to Figure 4, where we exemplify the various stages of ACE. For instance, we see that the pre-explanation contains out of distribution artifacts and how the refinement sends it back to the image distribution. Also, we highlight that the filtering modifies the hair, which is not an important trait for the classifier. The refinement is key to avoid editing these regions.

Effect of Different Adversarial Attacks. At the core of our optimization, we have the PGD attack. PGD is one of the most common attacks due to its strength. In this section, we explore the effect of incorporating other attacks. Thus, we tested C&W and the standard gradient descent (GD). Note that the difference between PGD and GD is that GD does not apply the $sign$ operation.

Our results show that these attacks are capable of generating semantic changes in the image. Although these are as successful as the PGD attack, we are required to optimize the pre-explanation for twice as many iterations. Even when our model is faster than [28], we still require about 250 DDPM iterations to generate a single explanation.

Validity of the S^3 Metric. In this paragraph, we show that the S^3 and the FS metrics are equivalent when used in the same test bed, *i.e.*, CelebA HQ. To this end, we assess whether the ordering between ACE, pre-explanation, and DiME are equal. To have a reference value, we evaluate the measurements when using a pair of random images. So, we show the values (ordering) for both metrics in Table 6 for the Age and Smiling attribute. As we expect, the ordering is similar between both metrics. Nevertheless, we stress that FS is adequate for faces since the network was trained for this task.

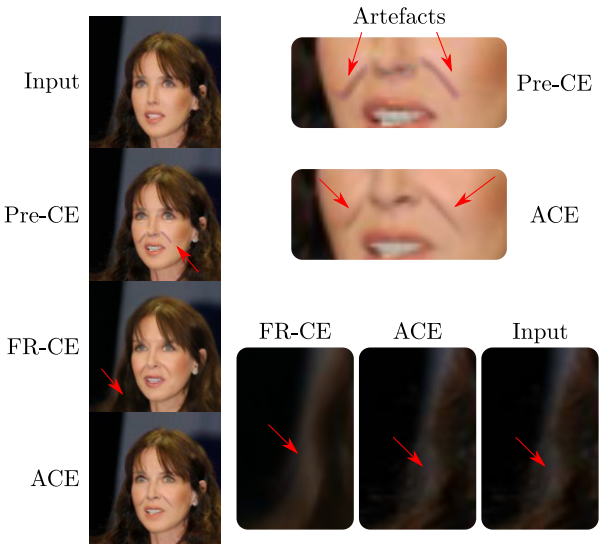

Figure 4. **Refinement Ablation.** We observe that pre-explanations can have out-of-distribution artifacts. After filtering them, the diffusion process creates in-distribution data, but there are unnecessary changes such as the background. ACE is capable of changing the key features while avoiding modifying unwanted structures.

Metric	Random	ACE	Pre-CE	DiME
Smile				
FS	0.2649 (4)	0.8941 (2)	0.9200 (1)	0.6729 (3)
S^3	0.4337 (4)	0.9876 (2)	0.9927 (1)	0.9396 (3)
Age				
FS	0.2649 (4)	0.7743 (2)	0.8300 (1)	0.6597 (3)
S^3	0.4337 (4)	0.9417 (2)	0.9870 (1)	0.9379 (3)

Table 6. **S^3 equivalence to FS.** The S^3 metric and the FS are equivalent in a similar context. We show the metric and the order (in parenthesis) and observe that both orderings are equal.

5. Conclusion

In this paper, we proposed ACE, an approach to generate counterfactual explanations using adversarial attacks. ACE relies on diffusion models to robustify the target classifier to create semantic changes via the adversary regardless of the classifier's robustness. ACE has multiple advantages regarding previous literature, notably seen in the counterfactual metrics. Moreover, we highlight that our explanations are capable of showing natural traits to find sparse and actionable modifications in real life, a feature not presented before. For instance, we were able to fool the classifier with physical modifications in the image as well as finding natural adversarial examples.

Acknowledgements: Research reported in this publication was supported by the Agence Nationale pour la Recherche (ANR) under award number ANR-19-CHIA-0017.

References

[1] Naveed Akhtar, Mohammad Jalwana, Mohammed Bennamoun, and Ajmal S Mian. Attack to fool and explain deep networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. 2020. 2

[3] Moritz Bohle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10029–10038, June 2021. 2

[4] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10329–10338, June 2022. 2

[5] Valentyn Boreiko, Maximilian Augustin, Francesco Croce, Philipp Berens, and Matthias Hein. Sparse visual counterfactual explanations in image space. In *DAGM German Conference on Pattern Recognition*, pages 133–148. Springer, 2022. 1, 2

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74, 2018. 4

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 2

[8] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 2

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, June 2021. 4

[10] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 12

[12] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2, 3, 5, 12

[13] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola Schoenlieb. On the connection between adversarial robustness and saliency map interpretability. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1823–1832. PMLR, 09–15 Jun 2019. 1

[14] Yunhao Ge, Yao Xiao, Zhi Xu, Meng Zheng, Srikrishna Karanam, Terrence Chen, Laurent Itti, and Ziyan Wu. A peek into the reasoning of neural networks: Interpreting with structural visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2195–2204, June 2021. 2

[15] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[16] 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. 2

[17] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384. PMLR, 09–15 Jun 2019. 2

[18] 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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 4

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 2, 3

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 4

[21] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[22] Frederik Hvilshøj, Alexandros Iosifidis, and Ira Assent. Ecinn: efficient counterfactuals from invertible neural networks. *British Machine Vision Conference 2018, BMVC 2021*, 2021. 2

[23] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. In H.

Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1, 2

[24] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 1

[25] Paul Jacob, Éloi Zablocki, Hedi Ben-Younes, Mickaël Chen, Patrick Pérez, and Matthieu Cord. STEEX: steering counterfactual explanations with semantics. In *ECCV*, 2022. 1, 4, 5

[26] Mohammad A. A. K. Jalwana, Naveed Akhtar, Mohammed Bennamoun, and Ajmal Mian. Cameras: Enhanced resolution and sanity preserving class activation mapping for image saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16327–16336, June 2021. 2

[27] Guillaume Jeanneret, Juan C. Pérez, and Pablo Arbeláez. A hierarchical assessment of adversarial severity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 61–70, October 2021. 2

[28] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, December 2022. 1, 2, 4, 5, 6, 8, 12

[29] Shalmali Joshi, Oluwasanmi Koyejo, Been Kim, and Joydeep Ghosh. xgems: Generating examplars to explain black-box models. *ArXiv*, abs/1806.08867, 2018. 4, 5

[30] Saeed Khorram and Li Fuxin. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10203–10212, June 2022. 2, 4

[31] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 10–15 Jul 2018. 2

[32] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. HIVE: Evaluating the human interpretability of visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[33] Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. Cartoon explanations of image classifiers. 2022. 2

[34] 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. 4, 5, 12

[35] Jeong Ryong Lee, Sewon Kim, Inyong Park, Taejoon Eo, and Dosik Hwang. Relevance-cam: Your model already knows where to look. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14944–14953, June 2021. 2

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 4, 5, 12

[37] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–665. Springer, 2021. 2

[38] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, June 2022. 4

[39] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2, 5

[40] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in artificial intelligence*, 5, 2022. 2

[41] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 2

[42] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 2

[43] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020. 4

[44] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Countergan: Generating realistic counterfactuals with residual generative adversarial nets. *arXiv preprint arXiv:2009.05199*, 2020. 2

[45] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022. 2

[46] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 4574–4594. PMLR, 2022. 2

[47] Juan C Pérez, Motasem Alfarra, Guillaume Jeanneret, Laura Rueda, Ali Thabet, Bernard Ghanem, and Pablo Arbeláez. Enhancing adversarial robustness via test-time transformation ensembling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[48] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018. 2

[49] Pau Rodríguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1056–1065, October 2021. 1, 2, 4, 5

[50] Dawid Rymarczyk, Łukasz Struski, Michał Górszczak, Koryna Lewandowska, Jacek Tabor, and Bartosz Zieliński. Interpretable image classification with differentiable prototypes assignment. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[51] Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[52] Lisa Schut, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Medb Corcoran, and Yarin Gal. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1756–1764. PMLR, 13–15 Apr 2021. 1, 2

[53] Sheng-Min Shih, Pin-Ju Tien, and Zohar Karnin. GANMEX: One-vs-one attributions using GAN-based model explainability. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9592–9602. PMLR, 2021. 2

[54] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. In *International Conference on Learning Representations*, 2020. 1, 2, 4, 5

[55] Sarah Tan, Rich Caruana, Giles Hooker, Paul Koch, and Albert Gordo. Learning global additive explanations for neural nets using model distillation, 2018. 2

[56] Jayaraman J. Thiagarajan, Vivek Narayanaswamy, Deepta Rajan, Jia Liang, Akshay Chaudhari, and Andreas Spanias. Designing counterfactual generators using deep model inversion. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2

[57] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 2

[58] Arnaud Van Looveren, Janis Klaise, Giovanni Vacanti, and Oliver Cobb. Conditional generative models for counterfactual explanations. *arXiv preprint arXiv:2101.10123*, 2021. 2

[59] Simon Vandenhende, Dhruv Mahajan, Filip Radenovic, and Deepti Ghadiyaram. Making heads or tails: Towards semantically consistent visual counterfactuals. In *ECCV 2022*, 2022. 2

[60] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *arvard Journal of Law and Technology,*, 31(2):841–887, 2018. 1

[61] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[62] Pei Wang, Yijun Li, Krishna Kumar Singh, Jingwan Lu, and Nuno Vasconcelos. Imagine: Image synthesis by image-guided model inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3681–3690, June 2021. 2

[63] Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8981–8990, 2020. 2

[64] 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. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9520–9529, 2020. 4, 5, 12

[65] 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. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642, 2020. 4, 5, 12

[66] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4

[67] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2

[68] Quan Zheng, Ziwei Wang, Jie Zhou, and Jiwen Lu. Shap-cam: Visual explanations for convolutional neural networks based on shapley value. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[69] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018. 2

[70] Yao Zhu, Jiacheng Ma, Jiacheng Sun, Zewei Chen, Rongxin Jiang, Yaowu Chen, and Zhenguo Li. Towards understanding the generative capability of adversarially robust classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7728–7737, 2021. 1

Supplementary Material: Adversarial Counterfactual Visual Explanations

A. Detailed Implementation Details

For each dataset, we used different configurations in architecture and for the generation of the pre-explanation. When using the distance loss ℓ_1, we set the distance regularization constant to $\lambda_d = 0.001$ while $\lambda_d = 0.1$ for ℓ_2. For the final refinement, firstly, we normalize the mask by the maximum pixel's difference magnitude. For the dilation step, we set the mask as a square with a width and height of 15 pixels for all datasets. Next, we will show all implementation details for each dataset.

CelebA [36]: We used the same architecture and weights as [28]. Additionally, we set $\tau = 5$ with a total amount of steps as 50. At the refinement stage, we used the same threshold of 0.15 for both ℓ_1 and ℓ_2 experiments for smile and age attributes.

CelebA HQ [34]: Our model follows the same architecture than [12] for ImageNet 256×256 unconditional generation. Since CelebA HQ is far less complex than ImageNet, we reduced the number of channels from 256 to 128. Also, our model generates samples using 500 diffusion steps instead of 1000. For training, we iterated our model for 120.000 iterations with a batch size of 256 on two V100 GPUs following [12]'s code. We set the learning rate to 10^4, a weight decay of 0.05, and no dropout.

To generate the pre-explanations, we noise the image until $\tau = 5$ out of 25 re-spaced steps. To binarize the mask, we used a threshold of 0.15 and 0.1 for the smiling attribute with the ℓ_1 and ℓ_2 distance losses, respectively. For the age attribute, we used 0.15 for ℓ_1 and 0.05 for ℓ_2.

BDD100k/OIA [64, 65]: The counterfactual explanation research community opted to use BDD100k in a 512×256 setup. This is highly demanding computationally to create a DDPM. Thus, since we knew *a priori* that we do not need many iterations for ACE to generate counterfactuals, we trained our diffusion model partially in the Markov chain. That is, our DDPM cannot generate images from pure noise. Instead, we trained it to generate images solely from a quarter of the complete chain, requiring an input instance to warm up the generation. So, we trained our model to generate instances with 250 steps out of 1000. This enabled us to use a lighter model. Artitecnologically, our UNet model has four downsampling stages with $128\,s$ channels, where s is the downsampling stage. Finally, we used the attention layer at the deeper layer of the UNet. At the training phase, we used a batch size of 256, a learning rate of 10^4, and a weight decay and dropout of 0.05 for 50.000 iterations.

To generate our explanations, we used 5 out of 100 (re-spaced) diffusion steps. For ℓ_1, we used a threshold of 0.05 and 0.1 for ℓ_2 for both datasets.

ImageNet [11]: For this dataset, we took advantage of previous works. In this case, we utilised [12]'s model on ImageNet 256. To generate the explanations, we used 5 steps out of 25 for the pre-explanations and set the threshold to 0.15 to binarize the mask for all cases.

B. Qualitative Results

In this section, we show more qualitative results. We will display the input image, its pre-explanation, the mask, and the final counterfactual for both ℓ_1 and ℓ_2 losses on all datasets. Note that we added a small discussion on the caption analyzing the results.

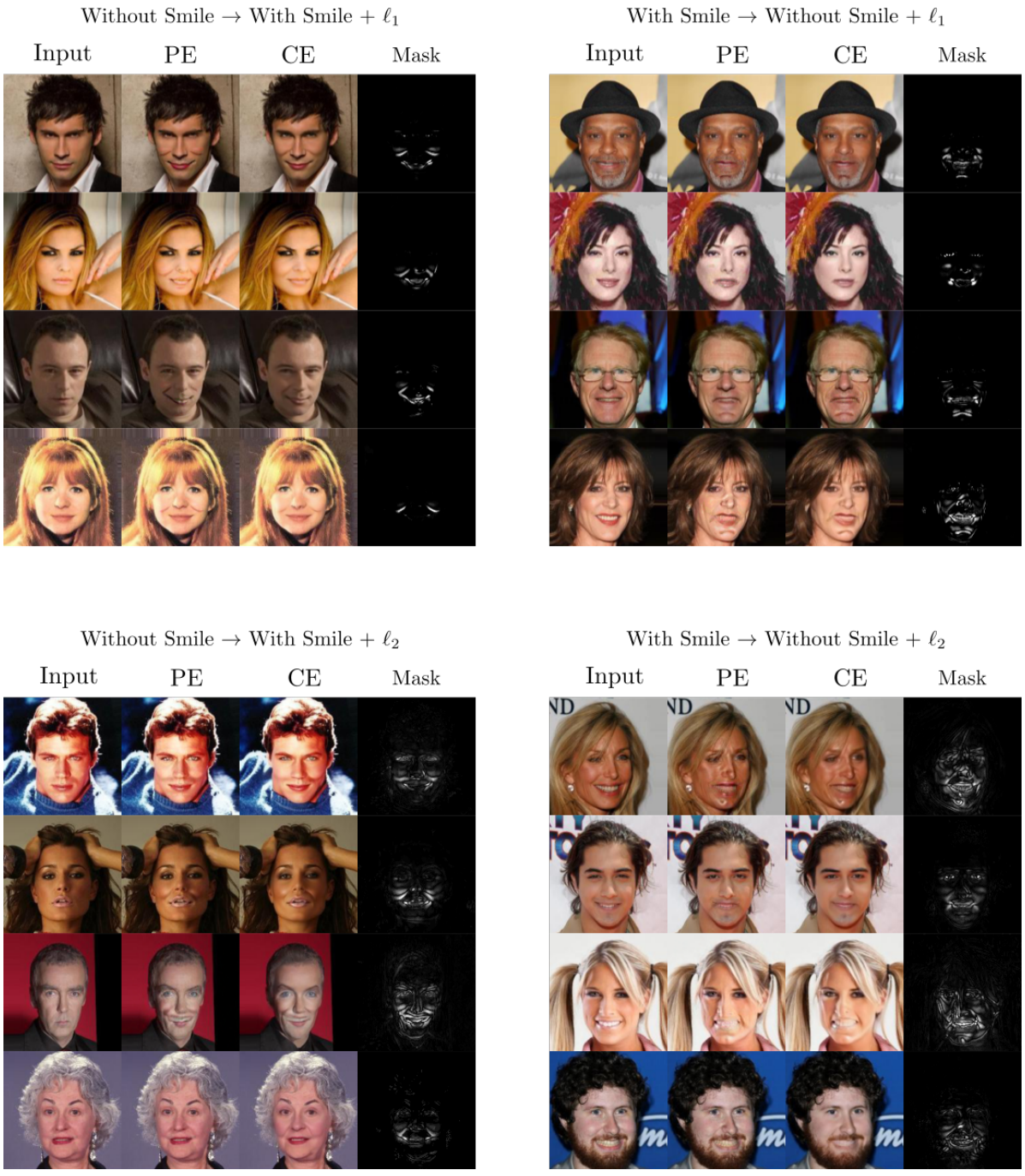

Figure 5. Additional CelebA qualitative results. We show examples for the *Smiling* attribute for both distances losses. From our qualitative experiments, we see that removing the smile attributes is harder than adding them. Additionally, we see that the ℓ_1 loss creates more sparse editings.

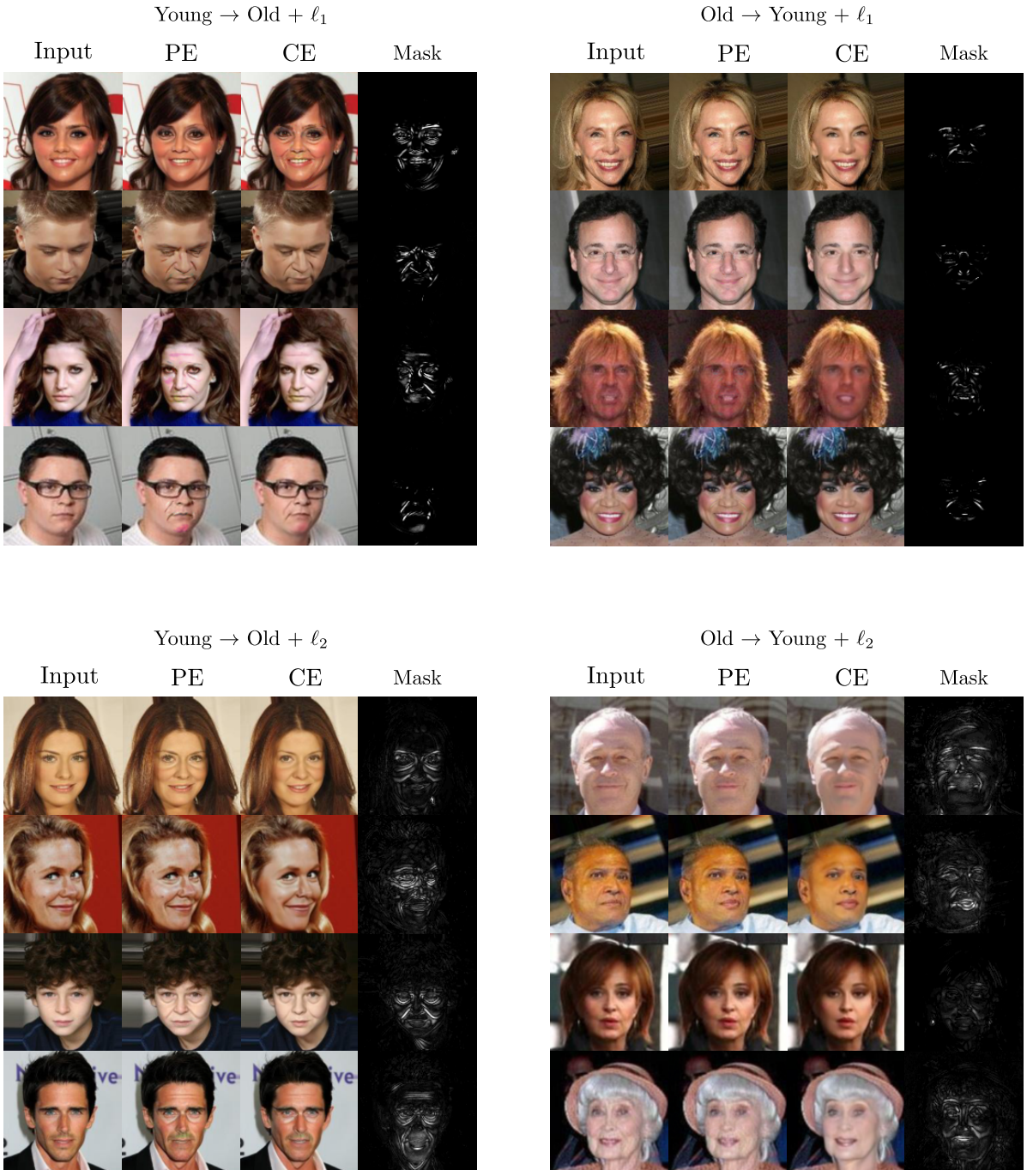

Figure 6. Additional CelebA qualitative results. We show examples for the *Age* attribute for both distances losses. The results show that the ℓ_1 loss creates more out-of-distribution artifacts.

Figure 7. Additional CelebA HQ qualitative results. We show examples for the *Smiling* attribute for both distances losses. We see similar behavior in the CelebA dataset.

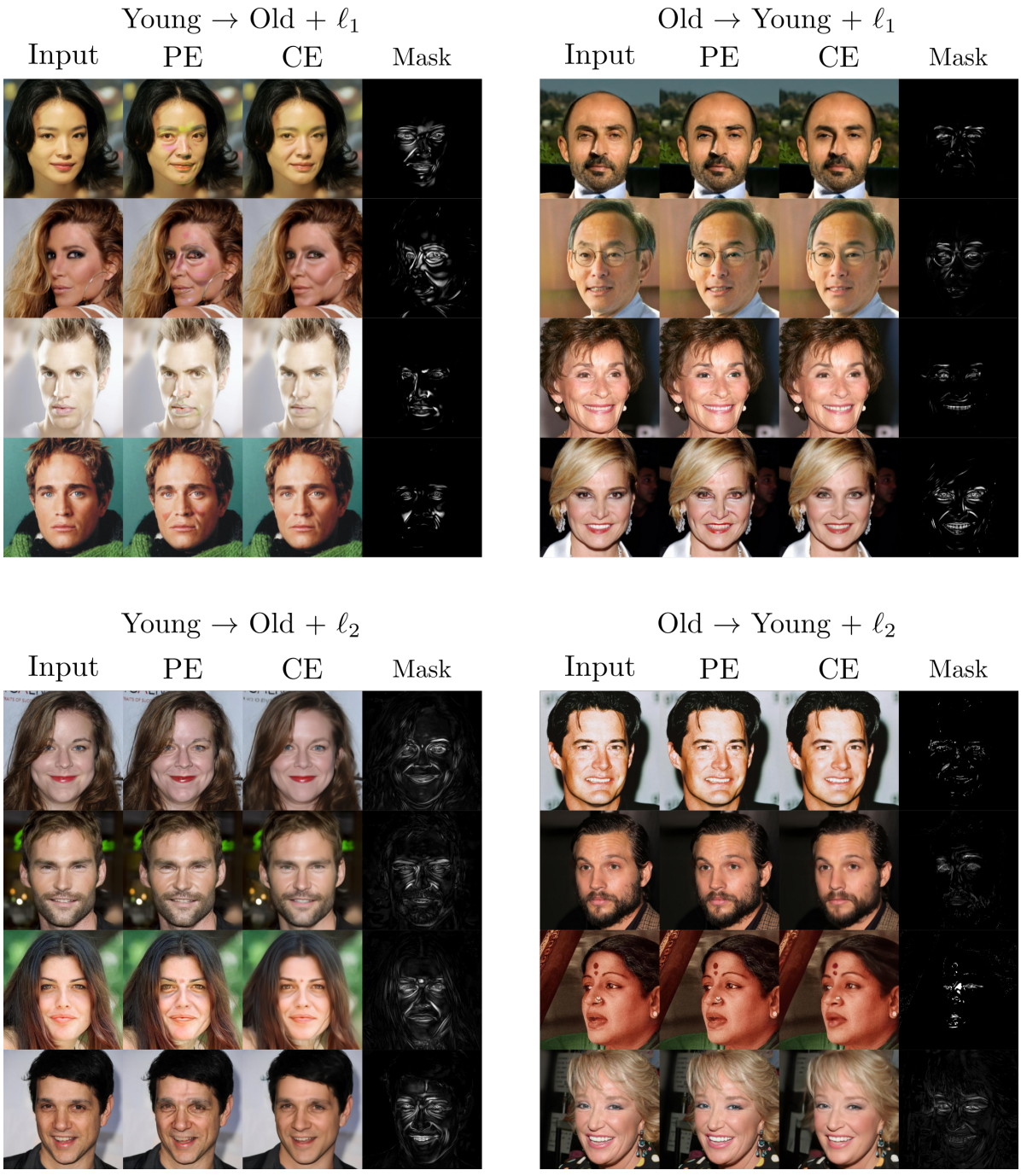

Figure 8. Additional CelebA HQ qualitative results. We show examples for the *Age* attribute for both distances losses. These examples show that transforming *Old* to *Young* is less informative than the other way.

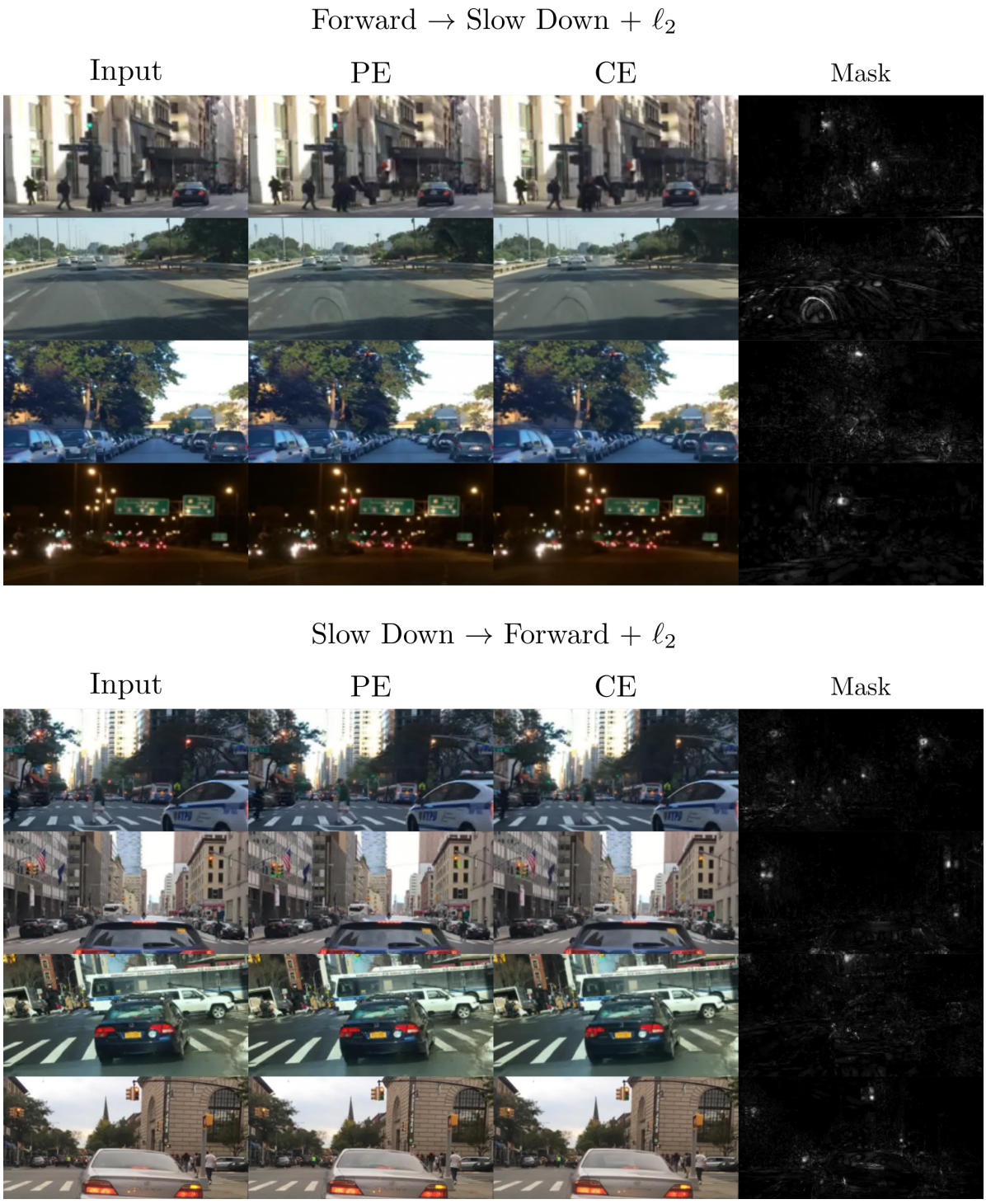

Figure 9. Additional BDD qualitative results. We show examples for the *Forward / Slow Down* binary class for ℓ_2 distance loss. We show a zoom of the changes in the image since the perturbations are sparse. We see that ACE adds traffic light colors in the buildings to change the prediction.

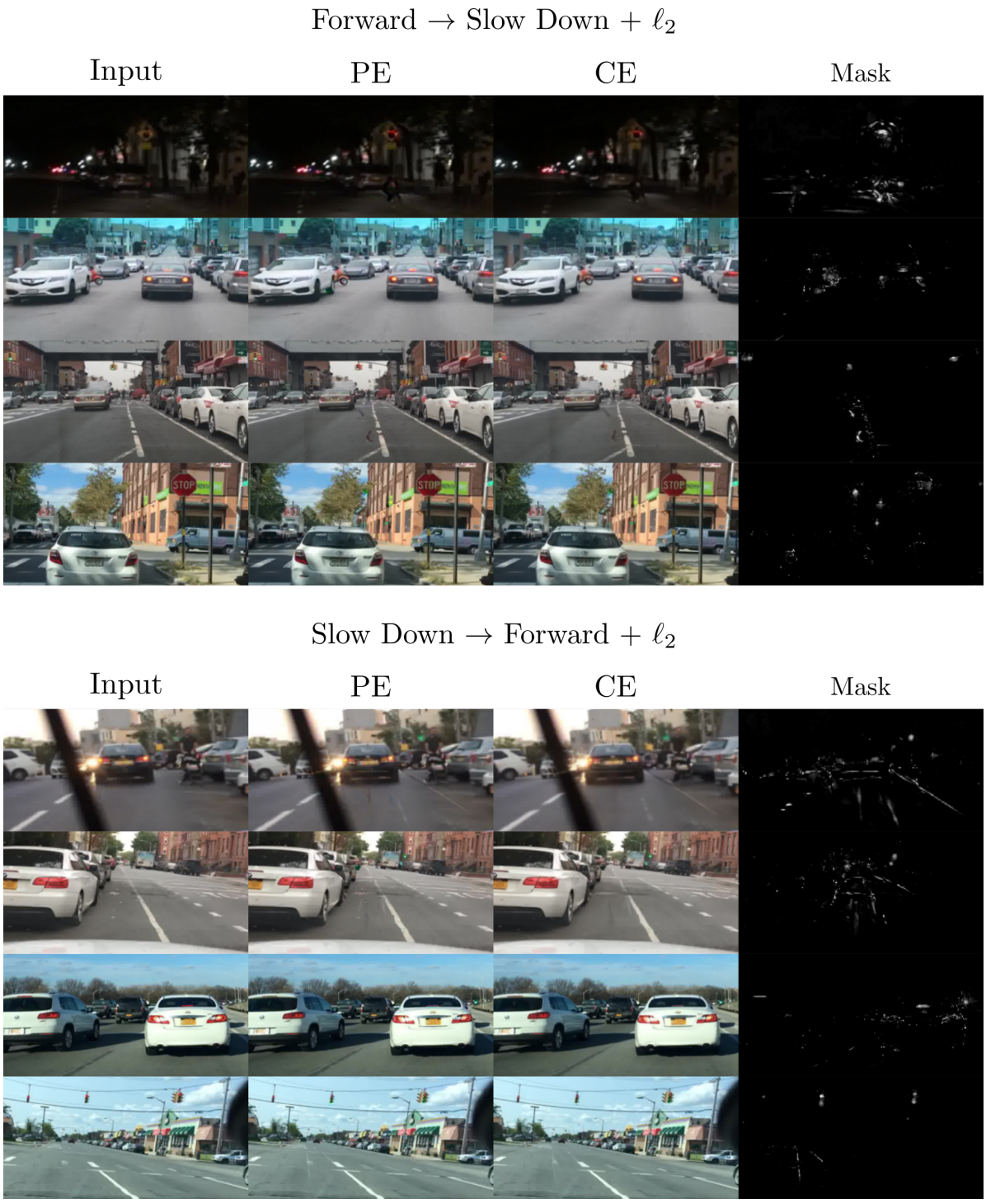

Figure 10. Additional BDD qualitative results. We show examples for the *Forward / Slow Down* binary class for ℓ_1 distance loss. We show a zoom of the changes in the image since the perturbations are sparse. We show a zoom of the changes in the image since the perturbations are sparse. We see that ACE adds traffic light colors in the buildings to change the prediction.

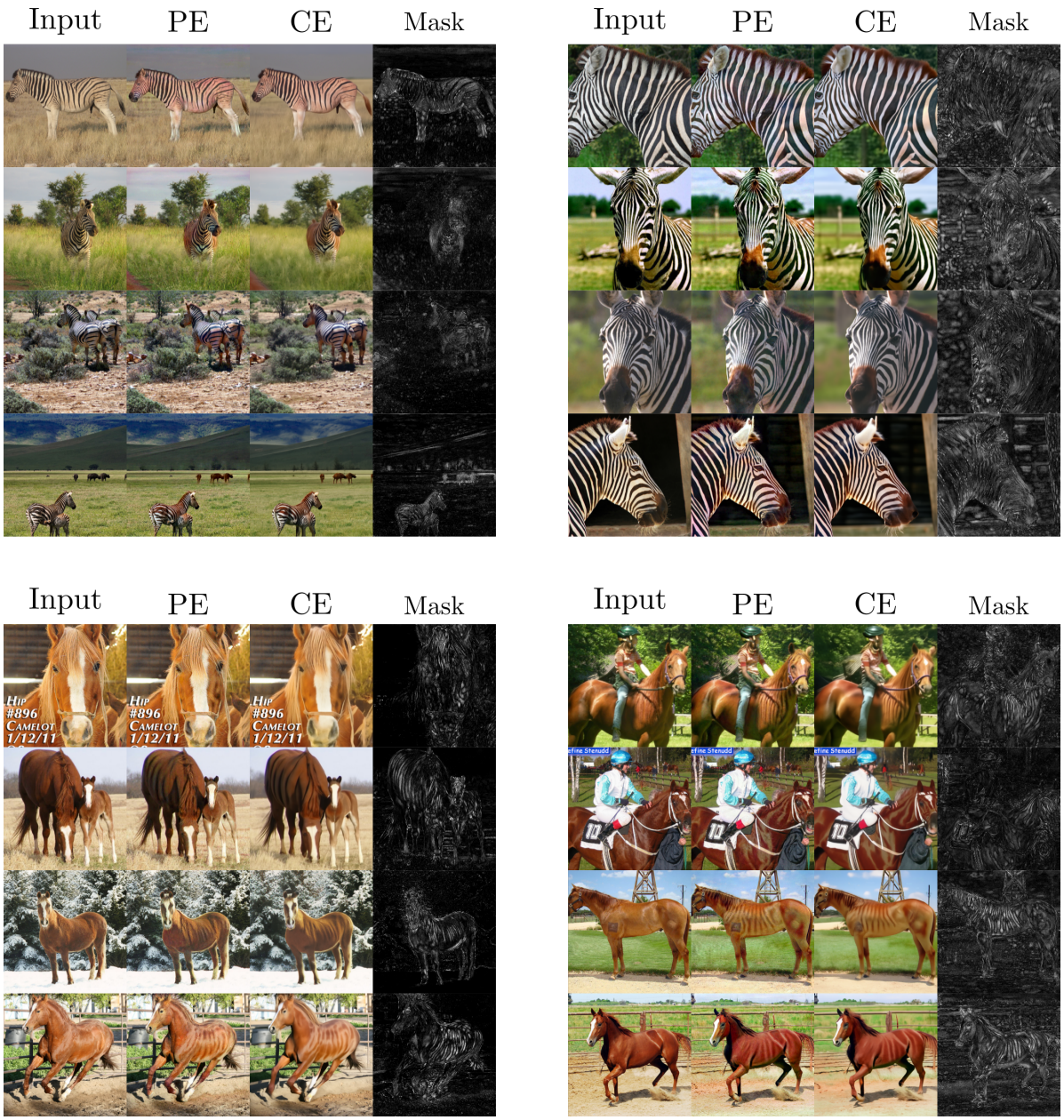

Figure 11. Additional ImageNet qualitative results. We show examples for the *Zebra / Sorrel* categories class. The first column is the ℓ_1 distance loss while the second one is ℓ_2. The initial row is zebra to sorrel and the second one is the inverse. To change from zebras to sorrels, some examples show not only incorporating the brown color sorrel horses but also the context in the background (*e.g.* adding a stable-like background). Vice-versa, to classify a horse as a zebra it is enough to add some strips.

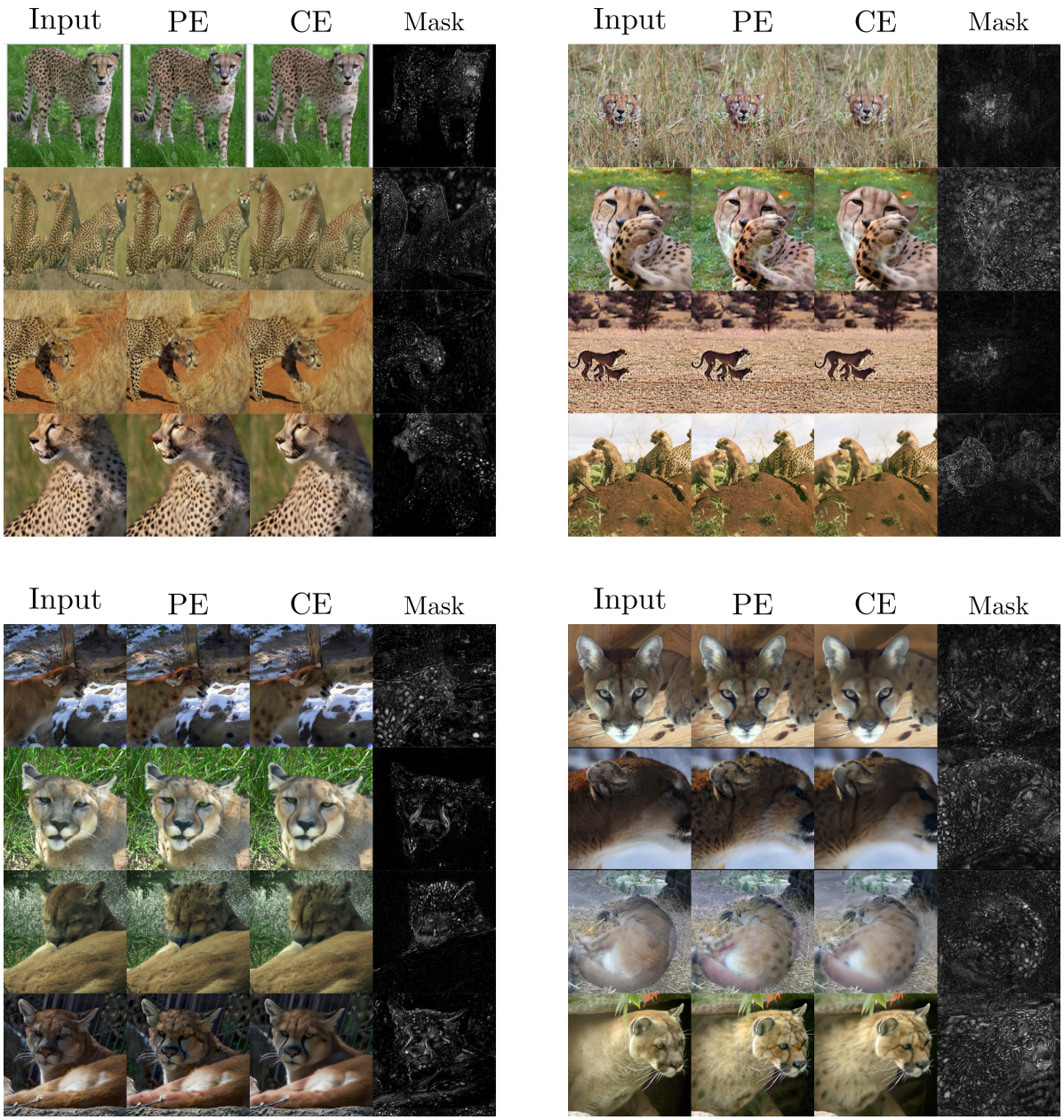

Figure 12. Additional ImageNet qualitative results. We show examples for the *Cheetah / Cougar* categories class. The first column is the ℓ_1 distance loss while the second one is ℓ_2. The first row is cheetah to cougar and the second is the inverse. We mainly see that changing from cheetah to cougar is enough to target the face of the animal. Vice-versa, to classify a cougar as a cheetah, ACE adds spots and characteristic cheetah stripes on the face.

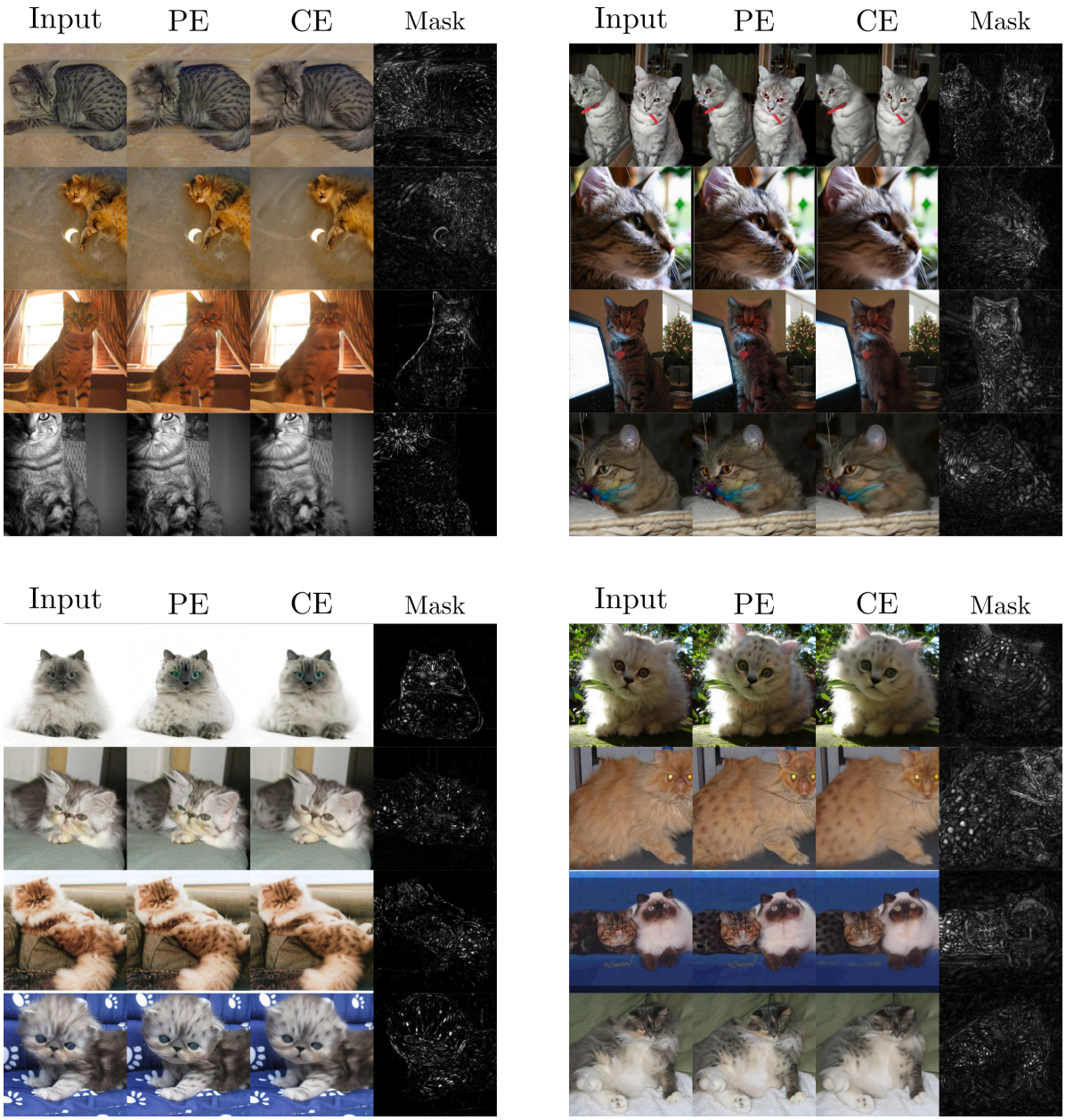

Figure 13. Additional ImageNet qualitative results. We show examples for the *Egyptian / Persian cat* categories class. The first column is the ℓ_1 distance loss while the second one is ℓ_2. The row is Egyptian to Persian cat and the second is the inverse. To change from Egyptian to Persian, we mainly see that ACE adds the Persian cats' fluffy fur. Conversely, from Persian to Egyptian it adds spots.