# Diffusion Visual Counterfactual Explanations

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Visual Counterfactual Explanations (VCEs) are an important tool to understand the decisions of an image classifier. They are "small" but "realistic" semantic changes of the image changing the classifier decision. Current approaches for the generation of VCEs are restricted to adversarially robust models and often contain non-realistic artefacts or are restricted to image classification problems with few classes. In this paper, we overcome this by generating Diffusion Visual Counterfactual Explanations (DVCEs) for arbitrary ImageNet classifiers via a diffusion process. Two modifications to the diffusion process are key for our DVCEs: i) an adaptive parameterization allows us to have hyperparameters that generalize across images and models and together with distance regularization and late start of the diffusion process allow us to generate close images and ii) our cone regularization via an adversarially robust model ensures that the diffusion process does not converge to trivial non-semantic changes but instead produces realistic images of the target class which achieve high confidence by the classifier.

## 1 Introduction

It can be argued that one of the main problems hindering the widespread use of machine learning is the missing possibility to explain the decisions of black-box models such as neural networks. This is not only a problem for decisions affecting humans where the current draft for AI regulation in Europe [9] requires "transparency", but it is a pressing problem in all applications of machine learning. The reason, to some extent, is that humans would like to understand but also control if the learning algorithm has captured the "concepts" of the underlying classes or if it just predicts well using spurious features, artefacts in the data set, or other sources of error. In this paper, we focus on model-agnostic explanations which can be applied to any classifier and do not rely on the specific structure of the classifier such as decision trees or linear classifiers. In this area, in particular for image classification, sensitivity based explanations [4], explanations based on feature attributions [3], saliency maps [40, 39, 14, 50, 45], Shapley additive explanations [26], and local fits of interpretable models [33] have been proposed.

Moreover, [49] proposed counterfactual explanations (CEs), which are instance-specific explanations. They can be applied to any classifier and it has been argued that they are close to human justification of decisions [28] using counterfactual reasoning: "I would have recognized it as zebra (instead of horse), if it would have had black and white stripes." For a given classifier almost all methods for the generation of CEs [49, 12, 29, 5, 32, 48, 37] try to solve the following problem: "Given a target class $c$ what is the minimal change $\delta$ of input $x$, such that $x + \delta$ is classified as class $c$ with high probability and is realistic instance of my data generating distribution?" From the perspective of "debugging" existing machine learning models, CEs are interesting as they construct a concrete input $x + \delta$ with different classification which allows the developer to test if the model has learned the correct features.

The main problem why visual counterfactual explanations (VCEs), that is CEs for image classification, are not widely used is that the problem of generating CEs is very related to the problem of adversarial

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Figure 1: Our method DVCE creates visual counterfactual explanations (VCEs) that explain any given classifier (here: non-robust ConvNeXt) by changing an image of one class into a similar target class by changing class relevant features in the foreground while preserving the overall image structure.

examples [46] and thus even imperceivable changes of the image can already change the prediction, which clearly does not show to the user if the classifier has picked up the right class-specific features. One possible solution are adversarially robust models [36, 1, 6], which have been shown to produce semantically meaningful VCEs by directly maximizing the probability of the target class in image space. These approaches have the downside that they only can generate VCEs for robust models which is a significant restriction as these models are not competitive in terms of prediction accuracy. The second approach is to restrict the generation of VCEs using a generative model or constraining the set of potential image manipulations [16, 17, 35, 7, 15, 38, 22]. However, these approaches are either restricted to datasets with a small number of classes, cannot provide explanations for arbitrary classifiers, or the generated images look realistic but are far away from the original one, so that the VCE and the original image have so little in common that not much insight can be gained.

Recently, [21] trained a StyleGAN model to discover and manipulate class attributes. While their approach yields impressive results on smaller-scale datasets with few similar classes (for example, different bird species), the authors did not demonstrate that the method scales to complex tasks such as ImageNet with hundreds of classes where different classes require different sets of attributes. Another disadvantage is that the StyleGAN model needs to be retrained for every classifier, making it prohibitively expensive to explain multiple large models.

In this paper, we overcome this by generating our Diffusion Visual Counterfactual Explanations (DVCEs) for arbitrary ImageNet classifiers. We use the progress in the generation of realistic images using diffusion processes [41, 20, 42, 44] reaching recently the level of GANs [11]. Similarly to [31, 2], we use a classifier and a distance-type regularization to guide the generation of the images. Two modifications to the diffusion process are key elements for our DVCEs: i) a combination of distance regularization and starting point of the diffusion process together with an adaptive reparameterization lets us generate VCEs visually close to the original image in a controlled way so that hyperparameters can be fixed across images and even across models, ii) our cone regularization of the gradient of the classifier via an adversarially robust model ensures that the diffusion process does not converge to trivial non-semantic changes but instead produces realistic images of the target class which achieve high confidence by the classifier. Our approach can be employed to any dataset where a generative diffusion model and an adversarially robust model are available. In a qualitative comparison and a user study in App. C we show that our DVCEs outperform the recent method of [6] as well as an adaptation of [2] for the generation of VCEs.

## 2 Diffusion models

Diffusion models [11, 41, 42, 20, 30] are generative models that consist of two steps: *forward diffusion* process (and thus a Markov process) that transforms the data distribution to a prior distribution (which is usually assumed to be a standard normal distribution) and the *reverse diffusion* process that transforms the prior distribution back to the data distribution.The existence of the (continuous-time) reverse diffusion process was closer investigated in [44].
In the discrete-time setting, a Markov chain $\{x_1, ..., x_T\}$ for any data point $x_0$ in the forward direction is defined via a Markovian process that adds noise to the data point at each timestep:

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{1}$$

where $\{\beta_1, ..., \beta_T\}$ is some variance schedule, chosen such that $q(x_T) \sim \mathcal{N}(0, I)$. Note that given $x_0$, it is possible to sample from $q(x_t|x_0)$ in closed form instead of applying noise $t$ times using:

$$q(x_t|x_0) = \sqrt{\overline{\alpha}_t} x_0 + \epsilon\sqrt{1 - \overline{\alpha}_t}, \epsilon \sim \mathcal{N}(0, I), \tag{2}$$

where $\overline{\alpha}_t := \prod_{k=1}^{t}(1 - \beta_k)$. As the number of total time steps $T$ approaches infinity, the reverse transitions $q(x_{t-1}|x_t)$ approach diagonal Gaussian distributions. Then one uses a DNN to predict mean $\mu_\theta$ and covariance $\Sigma_\theta$ for this reverse process with transition probabilities:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

To sample from $p_\theta(x_0)$, one samples $x_T$ from $\mathcal{N}(0, I)$ and then follows the reverse process by repeatedly sampling from the transition probabilities $p_\theta(x_{t-1}|x_t)$. Originally, diffusion models were trained by optimizing a variational lower bound of the data likelihood [41]. Later, [20] proposed a simplified loss $L_{simple}$ that resembles a denoising auto-encoder:

$$L_{\text{simple}}(\theta, T, p_{\text{data}}) := \sum_{t=1}^{T} \mathbb{E}_{x_0 \sim p_{\text{data}}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} ||\epsilon - \epsilon_\theta(x_t, t)||_2^2, \tag{4}$$

where $\mu$ can be recovered as: $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t))$. [30, 11] combined both losses to learn $\Sigma$ and $\mu$. In the continuous-time setting, [43, 19] have derived the respective lower bounds.

### 2.1 Class conditional sampling

There exists diffusion models trained with and without knowledge about classes in the dataset [11, 18]. For our experiments, we only use the class-unconditional diffusion model, such that all class-conditional features are introduced by the classifier which we want to explain. For the noise-aware classifier, the reverse process transitions are then of the form (Proof appendix H in [11]):

$$p_\theta(x_{t-1}|x_t, y) = Z\, p_\theta(x_{t-1}|x_t)\, p_\theta(y|x_t, t) \tag{5}$$

for a normalization constant $Z$. Typically $p_\theta(y|\cdot)$ is a time-step aware classifier specifically trained on noisy images belonging to time-step $t$ [11]. As further we would like to explain any classifier and not only noise-aware ones, we follow the approach of [2], where a classifier $p_\theta(y|\cdot)$ is given as an input the denoised sample $\hat{x}_0 = f_{\text{dn}}(x_t)$ of $x_t$, using the mapping:

$$f_{\text{dn}} : \mathbb{R}^d \times \{1, ..., T\} \to \mathbb{R}^d, (x_t, t) \mapsto \frac{x_t}{\sqrt{\overline{\alpha}_t}} - \frac{\sqrt{1 - \overline{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\overline{\alpha}_t}}. \tag{6}$$

This mapping is derived from eq. (2), from which we have that

$$x_0 = \frac{x_t}{\sqrt{\overline{\alpha}_t}} - \frac{\sqrt{1 - \overline{\alpha}_t}\epsilon}{\sqrt{\overline{\alpha}_t}}, \tag{7}$$

and because we use our diffusion model $\epsilon_\theta(x_t, t)$ to predict $\epsilon$. With this, we can define a time-step aware posterior $p_\theta(y|x_t, t)$ for any classifier $p_\theta(y|\cdot)$ as $p_\theta(y|x_t, t) = p_\theta(y|f_{\text{dn}}(x_t, t))$.

The issue is that efficient sampling from the original diffusion model is possible only because the reverse process is made of normal distributions. As we need to sample from $p_\theta(x_{t-1}|x_t, y)$ thousands of times to obtain a single sample from the data distribution, it is not possible to use MCMC-samplers with high complexity to sample from each of the individual transitions. In [11], they proposed to solve it by approximating $p_\theta(x_{t-1}|x_t, y)$ with slightly shifted versions of $p_\theta(x_{t-1}|x_t)$ to make closed-form sampling possible. Such transition kernels are given by:

$$p_\theta(x_{t-1}|x_t, y) = \mathcal{N}(\mu_t, \Sigma_\theta(x_t, t)), \tag{8}$$

where

$$\mu_t = \mu_\theta(x_t, t) + \Sigma_\theta(x_t, t)\nabla_{x_t} \log p(y|x_t). \tag{9}$$

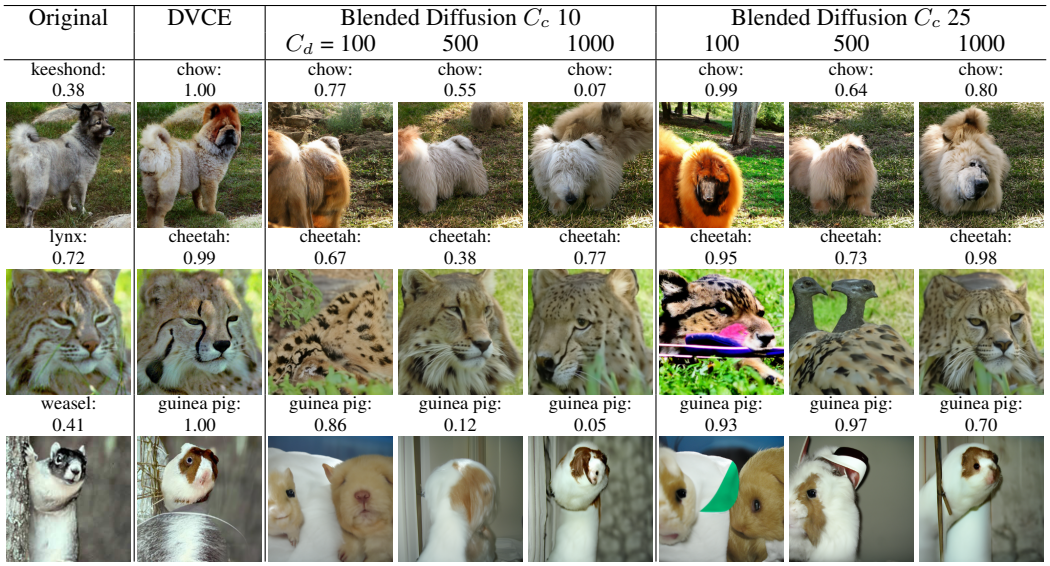


Figure 2: DVCEs for the robust Madry [13] + FT model (taken from [6]) in the second column, and DVCEs using the setting of [2], such that regularization is applied for the whole image. We can see that due to our adaptive parameterization, the same parameter works across different images for DVCEs ($C_c = 0.1, C_d = 0.15$), and across different classifiers (see Fig. 5). For blended diffusion, it is difficult to choose a set of hyperparameters that, even for a single classifier, works for multiple images. For the $l_2$ and LPIPS regularization weights of blended diffusion, we report the parameter of LPIPS and always use an $l_2$ weight 10 times as large (the ratio used in [2]).

## 3 Diffusion Visual Counterfactual Explanations

A VCE $z$ for a chosen target class $c$ and a given classifier $f$ and input $x$ should satisfy the following criteria: i) **validity:** the VCE $z$ should be classified by $f$ as the desired target class $c$ with high probability, ii) **realism:** the VCE $z$ should be indistinguishable or at least close to a natural image, iii) **minimality/closeness:** the difference of the VCE $z$ and the original image $x$ should be the minimal semantic modification necessary to change the class, in particular the generated image $z$ should be close to $x$ while being valid and realistic, e.g. by changing the object in the image and leaving the background unchanged. Note that while targeted adversarial examples are valid, they do not show meaningful semantic changes into the target class for a non-robust model and are also not realistic. Although adversarial $l_\infty$ perturbations are (almost) imperceivable for humans, adversarial images are very unlikely to occur in nature.

The $l_{1.5}$-SVCE method of [6] optimizes the predicted probability of classifier into the target class by manipulating the image and is very close to the generation of adversarial examples and thus works only for adversarially robust classifiers. The realism of the image comes purely from the generative properties of robust classifiers [36] which can lead to artefacts. In contrast, our Diffusion Visual Counterfactal Explanations (DVCEs) works for any classifier and our generated images are more realistic due to the integration of diffusion models. An approach similar to our DVCE framework is blended diffusion of [2] which manipulates the image inside a masked region. However, it is easy to adapt it to the task of VCEs by allowing the whole image to be manipulated. DVCE and blended diffusion share the same diffusion model, but, as we show in the appendix, blended diffusion cannot be applied to arbitrary classifiers and requires image-specific hyperparameter tuning, see Fig. 2, which is infeasible in practice.

### 3.1 Adaptive Parameterization

For the DVCE of the original image $\hat{x}$, we don't want to sample just any image $x$ from $p(x|y)$ (high realism and validity) but also to make sure that $d(x, \hat{x})$ is small in order to enforce a minimal change. To do this, we have to condition our diffusion process on $\hat{x}$. Using the introduced denoising function $f_{\text{dn}}$ and analogously to the derivation in [11] and following application in [22], the mean of the

transition kernel becomes:

$$\mu_\theta(x_t, t) + \Sigma_\theta(x_t, t)\nabla_{x_t}\Big[C_c \log p\big(y|f_{\text{dn}}(x_t, t)\big) - C_d d\big(\hat{x}, f_{\text{dn}}(x_t, t)\big)\Big], \tag{10}$$

where $C_c$ is the coefficient of the classifier, and $C_d$ the one of the distance guiding loss. Note that the last term can be interpreted as the log gradient of a distribution with density $\exp(-\frac{d_t(\hat{x},\cdot)}{\sigma})$, where $\sigma = \frac{1}{C_d}$ and $d_t(\hat{x}, \cdot) = d(\hat{x}, f_{\text{dn}}(\cdot, t))$, thus we introduce a time-aware prior distribution that enforces our output to be similar to $\hat{x}$. In our work, we use the $l_1$-distance as it produces sparser changes. Several works have tried to minimize some distance during the diffusion process, implicitly in [8] or explicitly, but for the background in [2]. However, there is no principled way to choose the coefficient for such a regularization term. In fact it turns out that it is impossible to find a parameter setting for blended diffusion [2] that works across images and classifiers. Thus, in this paper we propose an adaptive parametrization $g_{update}$

$$g_{\text{update}} = C_c \frac{\nabla_{x_t} \log p_\theta(y|f_{\text{dn}}(x_t, t)))}{\|\nabla_{x_t} \log p_\theta(y|f_{\text{dn}}(x_t, t)))\|_2} - C_d \frac{\nabla_{x_t} d(\hat{x}, f_{\text{dn}}(x_t, t))}{\|\nabla_{x_t} d(\hat{x}, f_{\text{dn}}(x_t, t))\|_2}, \tag{11}$$

relative to the predicted mean of the diffusion model that we use to change $\mu_t$ in Eq. 9 to

$$\mu_t = \mu_\theta(x_{t+1}, t+1) + \Sigma_\theta(x_{t+1}, t+1)\, \|\mu_\theta(x_{t+1}, t+1)\|_2\, g_{\text{update}}. \tag{12}$$

The adaptive parameterization allows for a finegrained control of the influcence of the classifier and distance regularization so that now the hyperparameters $C_c$ and $C_d$ have the same influence across images and even classifiers which facilitates the generation of DVCEs as otherwise hyperparameter finetuning would be necessary for each image as in blended diffusion, see Fig. 2 for a comparison. However, even with our adaptive parameterization, it is still not easy to produce semantically meaningful changes close to $\hat{x}$ as can be seen in App. B.2. Thus, as in [2], we vary the starting point in the diffusion process and observe in App. B.2 that starting from step $\frac{T}{2}$ of the forward diffusion process, together with the adaptive parameterization and using as distance the $l_1$-distance provides us with sparse but semantically meaningful changes.

### 3.2 Cone-Projection for Classifier Guidance

A major component of our DVCEs is that they can be applied to any image classifier regardless of whether it is adversarially robust or not. However, it turns out that naively applying diffusion with the new mean (Eq. 12) to a non-robust classifier does not work and leads to very small modifications of the image, which are similar to adversarial examples without meaningful changes. The reason is that gradients of non-robust classifiers are much more noisy and less semantically meaningful. We illustrate this for a non-robust Swin-Transformer[23] in Fig. 3 where hardly any changes are generated. Note that this is even the case if we denoise the sample first using the denoising function $f_{\text{dn}}$ (Eq. 6).

As a solution to this problem we suggest to project the gradient of a robust classifier, $\nabla_{x_t} \log p_{\text{robust},\theta}(y|f_{\text{dn}}(x_t, t))$, onto a cone centered at the gradient of the classifier, $\nabla_{x_t} \log p_\theta(y|f_{\text{dn}}(x_t, t))$, (we use $\alpha = 30°$ in all experiments for non-robust classifiers). More precisely, we define

$$g_{\text{proj}} = P_{\text{cone}(\alpha, \nabla_{x_t} \log p_\theta(y|f_{\text{dn}}(x_t,t)))}\Big[\nabla_{x_t} \log p_{\text{robust},\theta}\big(y|f_{\text{dn}}(x_t, t)\big)\Big],$$

where $\text{cone}(\alpha, v) := \{w \in \mathbb{R}^d : \angle(v, w) \leq \alpha\}$ is the cone of angle $\alpha$ around vector $v$, and the projection $P_{\text{cone}(\alpha,v)}$ onto $\text{cone}(\alpha, v)$ is given as:

$$P_{\text{cone}(\alpha,v)}[w] := \begin{cases} \langle u, w\rangle\, u & , \angle(w, v) > \alpha \\ w & , \text{else}, \end{cases}$$

where $P_{v^\perp}(w) := w - \frac{\langle w,v\rangle}{\langle v,v\rangle} v$ and we define $u = \sin(\alpha)\frac{P_{v^\perp}(w)}{\|P_{v^\perp}(w)\|_2} + \cos(\alpha)\frac{v}{\|v\|_2}$ (note that $\|u\|_2 = 1$ and $\langle u, v\rangle = \cos(\alpha)$). The idea of this projection is that in this way we reduce drastically the noise in the gradient of the non-robust classifier. Note, however, that the projection of the gradient of the robust classifier onto the cone generated by the non-robust classifier is always an ascent direction for $\log p_\theta(y|f_{\text{dn}}(x_t, t))$ and we still have the goal of maximizing this objective (note that $g_{\text{proj}}$ is not necessarily an ascent direction for $\log p_{\text{robust},\theta}(y|f_{\text{dn}}(x_t, t))$). Thus, one can see the cone projection as a form of regularization of $\nabla_{x_t} \log p_\theta(y|f_{\text{dn}}(x_t, t))$ which guides the diffusion process into semantically meaningful changes of the classifier.

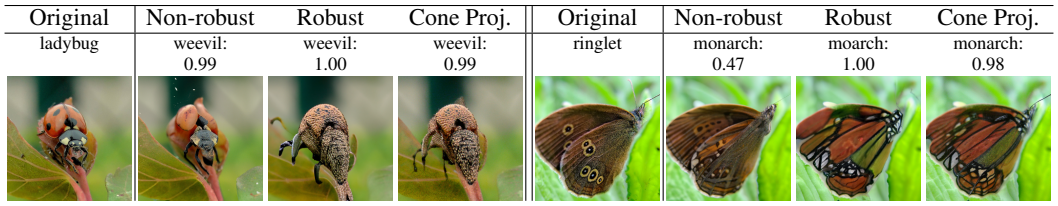


Figure 3: DVCEs for the non-robust Swin-Transformer[23], robust Madry[13] + FT model and DVCEs for the same non-robust model. While the adversarially robust classifier is guiding the diffusion process to produce class-specific features, a non-robust classifier does not yield perceptually aligned gradients. By using a cone projection of the robust and non-robust gradients, we can satisfy all desired properties of VCEs.

### 3.3 Final Scheme for Diffusion Visual Counterfactuals

For the convenience of the reader, we summarize here the full scheme of DVCEs for a non-adversarially robust classifier $p_\theta(y|x)$ where we assume that we are given additionally the adversarially robust model $p_{\text{robust},\theta}(y|x)$:

$$\begin{aligned}
g_{\text{proj}} &= P_{\text{cone}(\alpha, \nabla_{x_t} \log p_\theta(y|f_{\text{dn}}(x_t,t))}\left[\nabla_{x_t} \log p_{\text{robust},\theta}\big(y|f_{\text{dn}}(x_t,t)\big)\right], \\
g_{\text{update}} &= C_c \frac{g_{\text{proj}}}{\|g_{\text{proj}}\|_2} - C_d \frac{\nabla_{x_t} d(\hat{x}, f_{\text{dn}}(x_t,t))}{\|\nabla_{x_t} d(\hat{x}, f_{\text{dn}}(x_t,t))\|_2}, \\
\mu_t &= \mu_\theta(x_{t+1}, t+1) + \Sigma_\theta(x_{t+1}, t+1)\, \|\mu_\theta(x_{t+1}, t+1)\|_2\, g_{\text{update}}, \\
p(x_t|x_{t+1}, \hat{x}, y) &= \mathcal{N}(\mu_t, \Sigma_\theta(x_{t+1}, t+1)).
\end{aligned}$$

For a robust model the only difference is that the cone projection is omitted and one uses $g_{\text{proj}} = \nabla_{x_t} \log p_{\text{robust},\theta}\big(y|f_{\text{dn}}(x_t,t)\big)$. We use in all our experiments $C_c = 0.1$, and $C_d \in 0.15$ unless we show ablations for one of the parameters. The angle $\alpha$ for the cone projection is fixed to $30°$. In strong contrast to blended diffusion, these parameters generalize across images and even classifiers and allow us to generate VCEs for arbitrary classifiers for which a diffusion model is available.

## 4 Experiments

In this section we evaluate the quality of the visual counterfactual explanations of DVCE. We start by comparing DVCE to existing works in Sec. 4.1. In Sec. 4.2, we compare counterfactuals of DVCE for various state-of-the-art ImageNet models and show, how our method can be used to interpret differences between classifiers. For all experiments, we use the same set of hyperparameters for DVCE given in the last section. The diffusion model used for DVCE is taken from [11] and is trained class-unconditionally on ImageNet images of size 256x256 and uses a modified UNet architecture (more details in [11]). User study, further experiments, ablations and details can be found in the Appendix.

### 4.1 Method comparison

Creating VCEs for ImageNet with 1000 vastly different classes is a challenging task and there only exist a small number of methods that try to solve it without reducing dataset size or image resolution. For evaluation purposes, we compare our method to blended diffusion [2] and a recent method for VCEs based purely on optimization [6]. As the latter only works for adversarially robust classifiers, we use the multiple-norm robust model from [6], which they call Madry [13] + FT, as the classifier to create VCEs for all three methods. Results can be found in Fig. 4, where we transform one image into two different classes that are close in the WordNet hierarchy[27].

The radius of the $l_{1.5}$-ball of [6] is chosen as the smallest $r \in \{50, 75, 100, 150\}$ such that the confidence in the target class is larger than $0.90$ per image - if this is not reached, we show the one with highest confidence. For blended diffusion, we select the image with the smallest classifier and regularization weight that reaches a confidence larger than 0.9 from the set of parameters discussed in Sec. 3.3. If 0.9 is not reached by any setting, we again show the one with highest confidence. For

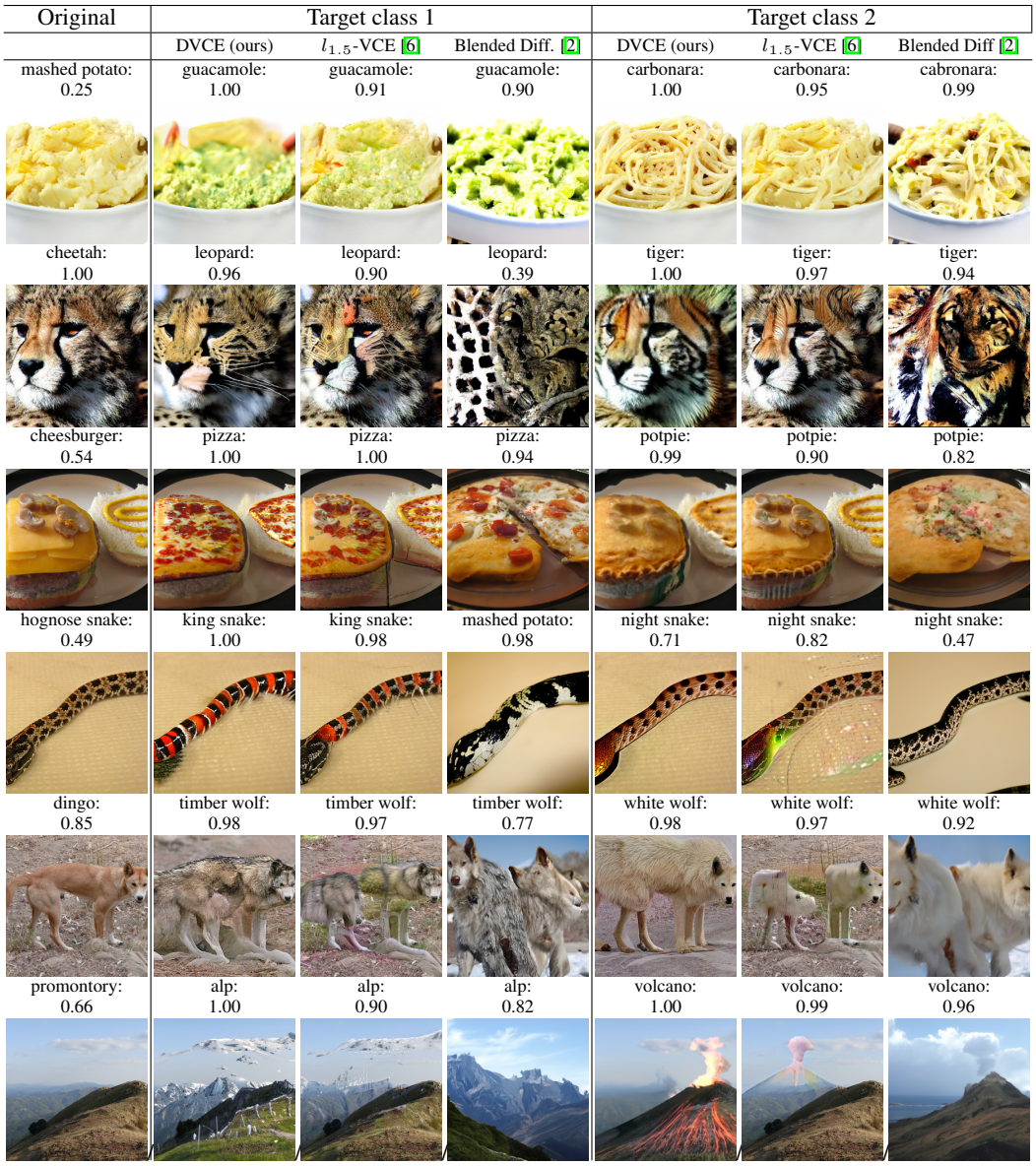


Figure 4: Comparison of different VCE-methods for the robust Madry[13] + FT (taken from [6]). We show our DVCEs (left column), the $l_{1.5}$-SVCEs of [6] (middle), and the adaptation of blended diffusion [2] for VCEs (right). For each original image (outer left column) we show the VCEs for two target classes. See the text for the parameter selection of all methods. Only DVCEs satisfy all desired properties of VCEs. Blended diffusion fails completely for leopard and tiger and often produces images far away from the original one (pizza, potpie, timber wolf, white wolf and alp). $l_{1.5}$-SVCEs show artefacts (white wolf, volcano, night snake) and have often lower image quality.

DVCEs, we only use the single set of parameters provided in the previous section. As the model is robust its own, we do not use the cone projection in DVCE.

As Figure 4 shows, DVCEs is the only method that satisfies all desired properties of VCEs: i) the produced VCEs have high confidence in the target class, ii) are close to natural images and iii) show minimal semantic changes into the target class. For example, for "mashed potato", DVCE preserves the bowl and only changes the content into either guacamole or carbonara. Our qualitative comparison shows that the same hyperparameter setting of DVCE can handle a variety of classes and transfer between similar classes, such as different snakes or wolf types, as well as different object sizes. For example, it works for "cheetah" which covers the entire image, but also for the "snake" which only

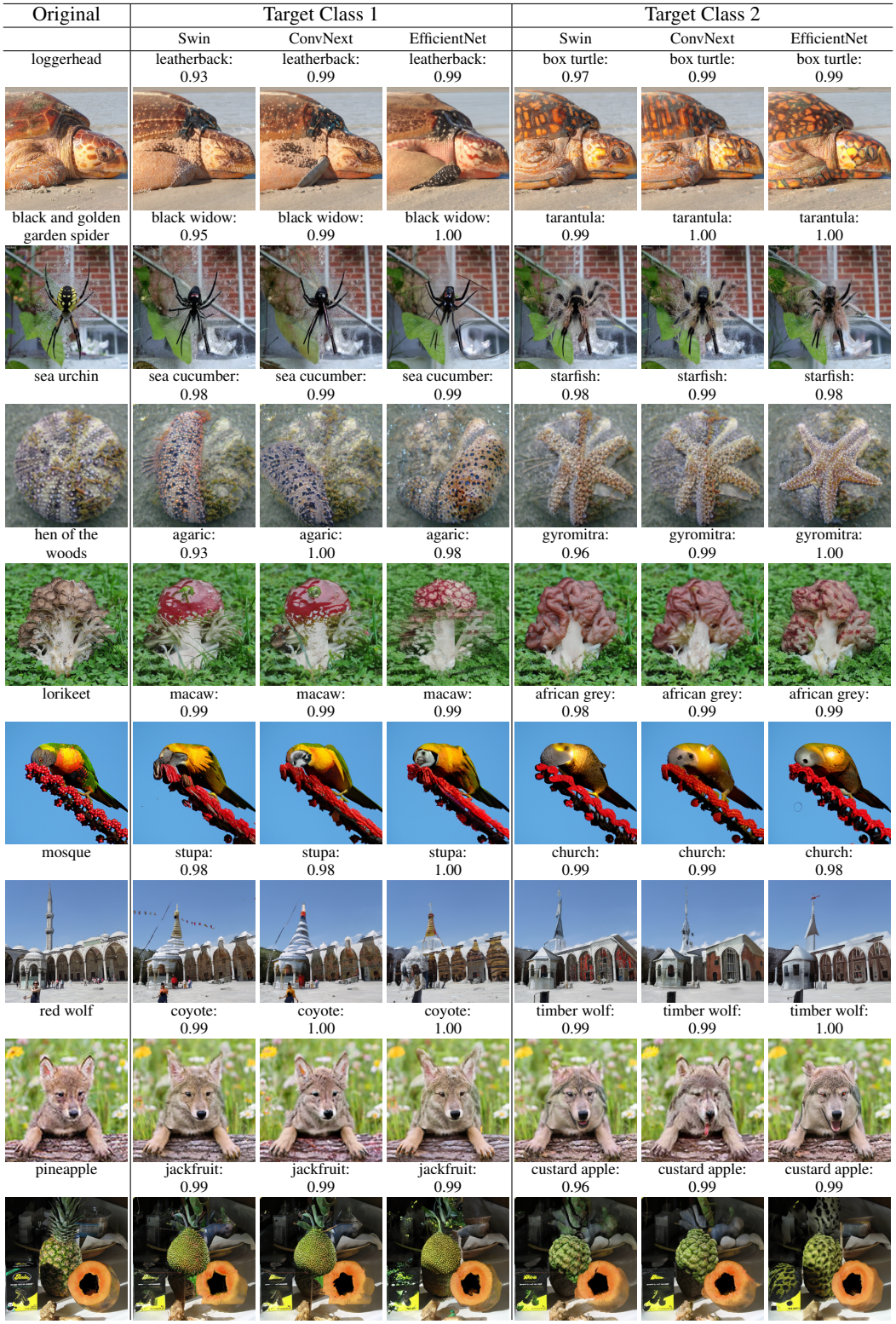


Figure 5: DVCEs for three non-robust classifiers (with ImageNet accuracy): Swin-T Transformer [23] (86.31%), ConvNext (87.02%) [24] and EfficientNet[51] (86.83%). For each original image (outerleft column) we show DVCEs into two target classes. Please zoom into the images to see more finegrained details and differences between the generated images.

covers a fraction of the input. In contrast, both $l_{1.5}$-SVCEs and blended diffusion VCEs require different hyperparameters for different images to achieve high confidence into the target class. Even with the six parameter configurations, blended diffusion is not able to produce always images with high confidence in the target class. Even more problematic for VCEs: sometimes the resulting images have high confidence but do not look realistic (cheetah $\rightarrow$ tiger) or do not resemble the original image at all (dingo $\rightarrow$ timber wolf/white wolf). Even if the method works with the given parameters, for example mashed potato $\rightarrow$ guacamole or carbonara, the overall image quality cannot match that of DVCE as often the images contain overly bright colors. In the case of the volcano VCE, DVCE shows class features like lava whereas blended diffusion VCE can not clearly be labeled as a volcano. For the $l_{1.5}$-SVCEs, one often needs large radii to achieve the desired confidence of 0.90, which often results in color artefacts and images that generally do not look realistic.

### 4.2 Model comparison

In Fig. 5 we show that DVCEs can be generated for various state-of-the-art ImageNet [34] models. We use a Swin-Transformer[23], a ConvNeXt [24] and a Noisy-Student EfficientNet [51, 47]. Both the Swin-Tranformer and the ConvNeXt are pretrained on ImageNet21k [10] wheareas the EfficientNet uses noisy-student self-training on a large unlabeled pool. As all models do not yield perceptually aligned gradients, we use the cone projection with $30^\circ$ angle described in Sec. 3.2 with the robust model from the previous section. All other parameters are identical to the previous experiment. We highlight that DVCE is the first method capable of visualizing arbitrary classifiers on a task as challenging as ImageNet [34]. Overall, DVCEs satisfy all desired properties of VCEs. It also allows us to inspect the most important features for each model and class. For example, all three models have learnt to associate the specific texture pattern of the box turtle with that particular class. For the stupa and church classes, it seems like all models use different roof and tower structures as the most prominent feature as they spent most of their budget on changing those.

## 5 Limitations, Future Work and Societal impact

In comparison to GANs, diffusion-based approaches can be expensive to evaluate as they require multiple iterations of the reverse process to create one sample. This can be an issue when creating a large amount of VCEs and makes deployment in a time-sensitive setting challenging.

We also rely on a robust model during the creation of VCEs for the cone projection. While we show that standard classifiers do not yield the desired gradients, training robust models can be challenging. An interesting direction for future research is thus to replace the cone projection with another "denoising" procedures for the gradient of a non-robust model.

From a theoretical standpoint, using conditional sampling with reverse transitions of the form Eq. 5 is justified, in practice, however, we have to approximate the reverse transitions with shifted normal distributions of the form Eq. 8. This approximation relies on the conditioning function having low curvature, and it can be hard to verify this once we start adding a classifier, distance and other possible terms and it is unclear how this influences the outcome of the diffusion process.

The quantitative evaluation of VCEs is difficult, as the standard FID metric compares the distribution of features of a classifier over a test and a generated dataset. However, for VCEs it is not only important to generate realistic images, but also achieve a high confidence and to create meaningful changes. Moreover, metrics such as IM1, IM2 [25] for VCEs rely on a well-trained (V)AE for every class, which is difficult to achieve for a dataset with 1000 classes and high-resolution images. Future research should therefore try to develop metrics for the quantitative evaluation of VCEs.

DVCEs, and VCEs in general help to discover biases of the classifiers and thus have a positive societal impact, however one can abuse them for unintended purposes as any conditional generative model.

## 6 Conclusion

We have proposed DVCEs, a novel way to create VCEs for any state-of-the-art ImageNet classifier using diffusion models and our cone projection. Our DVCEs satisfy all desired properties of VCEs. Throughout the paper, we show that DVCEs can handle vastly different image configurations, object sizes and classes.

# References


[1] Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in- and out-distribution improves explainability. In *ECCV*, 2020.

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022.

[3] Sebastian Bach, Alexander Binder, Frederick Klauschen Gregoire Montavon, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, 2015.

[4] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *JMLR*, 2010.

[5] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *FAT*, 2020.

[6] Valentyn Boreiko, Maximilian Augustin, Francesco Croce, Philipp Berens, and Matthias Hein. Sparse visual counterfactual explanations in image space. *arXiv preprint*, 2022.

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

[8] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *ICCV*, 2021.

[9] EU Commission. Regulation for laying down harmonised rules on AI. *European Commission*, 2021.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: a large-scale hierarchical image database. *CVPR09*, 2009. License: No license specified.

[11] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

[12] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *NeurIPS*, 2018.

[13] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. License: MIT.

[14] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. In *ICML*, 2019.

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

[16] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *ECCV*, 2016.

[17] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *ECCV*, 2018.

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[19] Chin-Wei Huang, Jae Hyun Lim, and Aaron Courville. A variational perspective on diffusion-based generative models and score matching. In *NeurIPS*, 2021.

[20] Pieter Abbeel Jonathan Ho, Ajay Jain. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[21] 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. *arXiv preprint*, 2021.

[22] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *CoRR*, 2021.

[23] Ze Liu, Fand Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[24] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint*, 2022.

[25] Arnaud Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint*, 2019.

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

[27] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 1995.

[28] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2019.

[29] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAccT*, 2020.

[30] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint*, 2021.

[32] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *NeurIPS*, 2020.

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

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015. License: No license specified.

[35] Pouya Samangouei, Ardavan Saeedi, Liam Nakagawa, and Nathan Silberman. Explaingan: Model explanation via decision boundary crossing transformations. In *ECCV*, 2018.

[36] Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *NeurIPS*, 2019.

[37] Lisa Schut, Oscar Key, Rory McGrath, Luca Costabello, Bogdan Sacaleanu, Medb Corcoran, and Yarin Gal. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *AISTATS*, 2021.

[38] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using stylegan for visual interpretability of deep learning models on medical images. In *NeurIPS Workshop "Medical Imaging Meets NeurIPS"*, 2020.

[39] 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. *International Journal of Computer Vision*, 2019.

[40] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2014.

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[43] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *NeurIPS*, 2021.

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[45] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *NeurIPS*, 2019.

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

[47] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

[48] Sahil Verma, John P. Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint*, 2020.

[49] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 2018.

[50] Zifan Wang, Haofan Wang, Shakul Ramkumar, Matt Fredrikson, Piotr Mardziel, and Anupam Datta. Smoothed geometry for robust attribution. In *NeurIPS*, 2020.

[51] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020.

# Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sec. 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Sec. 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] We didn't train models.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] We have conducted qualitative experiments in the main paper for the same seed and show the diversity of VCEs across seeds in App. B.3.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See App. D.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [Yes] Yes, directly in citation, when applicable.

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include our code in supplemental material.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] ImangeNet and ImageNet21k are public datasets.

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] We include the results from the user study together with the screenshots of instructions in App. C.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Not applicable.

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See App. C.