Beyond Trivial Counterfactual Explanations with Diverse Valuable Explanations

Pau Rodríguez[1] Massimo Caccia[1,2,3] Alexandre Lacoste[1] Lee Zamparo[1] Issam Laradji[1,2,4]
Laurent Charlin[2,5,6] David Vazquez[1]
[1]Element AI [2]MILA [3]UdeM [4]McGill University [5]HEC Montreal [6]CIFAR AI Chair
pau.rodriguez@servicenow.com

Abstract

Explainability for machine learning models has gained considerable attention within the research community given the importance of deploying more reliable machine-learning systems. In computer vision applications, generative counterfactual methods indicate how to perturb a model's input to change its prediction, providing details about the model's decision-making. Current methods tend to generate trivial counterfactuals about a model's decisions, as they often suggest to exaggerate or remove the presence of the attribute being classified. For the machine learning practitioner, these types of counterfactuals offer little value, since they provide no new information about undesired model or data biases. In this work, we identify the problem of trivial counterfactual generation and we propose DiVE to alleviate it. DiVE learns a perturbation in a disentangled latent space that is constrained using a diversity-enforcing loss to uncover multiple valuable explanations about the model's prediction. Further, we introduce a mechanism to prevent the model from producing trivial explanations. Experiments on CelebA and Synbols demonstrate that our model improves the success rate of producing high-quality valuable explanations when compared to previous state-of-the-art methods. Code is available at https://github.com/ElementAI/beyond-trivial-explanations.

1. Introduction

Consider an image recognition model such as a smile classifier. In case of erroneous prediction, an explainability system should provide information to machine learning practitioners to understand why such error happened and how to prevent it. Counterfactual explanation methods [3, 8, 9] can help highlight the limitations of an ML model by uncovering data and model biases. Counterfactual explanations provide perturbed versions of the input data that emphasize features that contributed the most to the ML model's output. For the smile classifier, if the model is confused by people wearing sunglasses then the system could generate alternative images of faces without sunglasses that would be correctly recognized. In order to discover a model's limitations, counterfactual generation systems could be used to generate images that would confuse the classifier, such as people wearing sunglasses or scarfs occluding the mouth. This is different from other types of explainability methods such as feature importance methods [3, 39, 40] and boundary approximation methods [29, 36], which highlight salient regions of the input like the sunglasses but do not indicate how the ML model could achieve a different prediction.

According to [30, 37], counterfactual explanations should be *valid*, *proximal*, and *sparse*. A *valid* counterfactual explanation changes the prediction of the ML model, for instance, adding sunglasses to confuse a smile classifier. The explanation is *sparse* if it only changes a minimal set of attributes, for instance, it only adds sunglasses and it does not add a hat, a beard, or the like. An explanation is *proximal* if it is perceptually similar to the original image, for instance, a ninety degree rotation of an image would be a sparse but not proximal. In addition to the three former properties, generating a set of *diverse* explanations increases the likelihood of finding a useful explanation [30, 37]. A set of counterfactuals is diverse if each one proposes to change a different set of attributes. Following the previous example, a diverse set of explanations would suggest to add or remove sunglasses, beard, or scarf, while a non-diverse set would all suggest to add or remove different brands of sunglasses. Intuitively, each explanation should shed light on a different action that a user can take to change the ML model's outcome.

Current generative counterfactual methods like xGEM [18] generate a single explanation that is not constrained to be *similar* to the input. Thus, they fail to be *proximal*, *sparse*, and *diverse*. Progressive Exaggeration (PE) [42] provides higher-quality explanations that are more *proximal* than xGEM, but it still fails to provide a *diverse* set of explanations. In addition, the image generator of PE is trained on the same data as the image classifier in

order to detect biases thereby limiting their applicability. Both of these two methods tend to produce *trivial* explanations, which only address the attribute that was intended to be classified, without further exploring failure cases due to biases in the data or spurious correlations. For instance, an explanation that suggests to increase the 'smile' attribute of a 'smile' classifier for an already-smiling face is trivial and it does not explain why a misclassification occurred. On the other hand, a non-trivial explanation that suggests to change the facial skin color would uncover a racial bias in the data that should be addressed by the ML practitioner. In this work, we focus on *diverse valuable* explanations, that is, *valid*, *proximal*, *sparse*, and *non-trivial*.

We propose Diverse Valuable Explanations (DiVE), an explainability method that can interpret ML model predictions by identifying sets of valuable attributes that have the most effect on model output. In order to generate *non-trivial* explanations, DiVE leverages the Fisher information matrix of its latent space to focus its search on the less influential factors of variation of the ML model. This mechanism enables the discovery of spurious correlations learned by the ML model. DiVE produces multiple counterfactual explanations which are enforced to be *valuable*, and *diverse*, resulting in more informative explanations for machine learning practitioners than competing methods in the literature. Our method first learns a generative model of the data using a β-TCVAE [4] to obtain a disentangled latent representation which leads to more *proximal* and *sparse* explanations. In addition, the VAE is not required to be trained on the same dataset as the ML model to be explained. DiVE then learns a latent perturbation using constraints to enforce *diversity*, *sparsity*, and *proximity*.

We provide experiments to quantify the success of explainability systems at finding *valuable* explanations. We find that DiVE is more successful at finding *non-trivial* explanations than previous methods and baselines. In addition, we provide experiments to compare the quality of the generated explanations with the current state-of-the-art. First, we assess their *validity* on the CelebA dataset [26] and provide quantitative and qualitative results on a bias detection benchmark [42]. Second, we show that the generated explanations are more *proximal* in terms of Fréchet Inception Distance (FID) [14], which is a measure of similarity between two datasets of images commonly used to evaluate the quality of generated images. In addition, we evaluate the *proximity* in latent space and face verification accuracy, as reported by Singla et al. [42]. Third, we assess the *sparsity* of the generated counterfactuals by computing the average change in facial attributes.

We summarize the contributions of this work as follows: 1) We identify the importance of finding *non-trivial* explanations and we propose a new benchmark to evaluate how *valuable* the explanations are. 2) We propose DiVE, an explainability method that can interpret an ML model by identifying the attributes that have the most effect on its output. 3) We propose to leverage the Fisher information matrix of the latent space for finding spurious features that produce *non-trivial* explanations. 4) DiVE achieves state of the art in terms of the *validity*, *proximity*, and *sparsity* of its explanations, detecting biases on the datasets, and producing multiple explanations for an image.

2. Related Work

Explainable artificial intelligence (XAI) is a suite of techniques developed to make either the construction or interpretation of model decisions more accessible and meaningful. Broadly speaking, there are two branches of work in XAI, ad-hoc and post-hoc. Ad-hoc methods focus on making models interpretable, by imbuing model components or parameters with interpretations that are rooted in the data themselves [17, 31, 35]. To date, most successful machine learning methods, including deep learning ones, are uninterpretable [5, 12, 16, 24].

Post-hoc methods aim to explain the decisions of uninterpretable models. These methods can be categorized as non-generative and generative. Non-generative methods use information from an ML model to identify the features most responsible for an outcome for a given input. Approaches like [29, 32, 36] interpret ML model decisions by fitting a locally interpretable model. Others use the gradient of the ML model parameters to perform feature attribution [1, 39–41, 43, 47, 48], sometimes by employing a reference distribution for the features [8, 40]. This has the advantage of identifying alternative feature values that, when substituted for the observed values, would result in a different outcome. These methods are limited to small contiguous regions of features with high influence on the target model outcome. In so doing, they can struggle to provide plausible changes of the input that are useful for a user in order to correct a certain output or bias of the model. Generative methods such as [3, 4, 6, 11] propose *proximal* modifications of the input that change the model decision. However the generated perturbations are usually performed in pixel space and bound to masking small regions of the image without necessarily having a semantic meaning. Closest to our work are generative counterfactual explanation methods which synthesize perturbed versions of observed data that result in a change of the model prediction. These can be further subdivided into two families. The first family of methods conditions the generative model on attributes, by e.g. using a conditional GAN [18, 25, 38, 45, 46]. This dependency on attribute information can restrict the applicability of these methods in scenarios where annotations are scarce. Methods in the second family use generative models such as VAEs [20] or unconditional GANs [10] that do not depend on attributes during generation [7, 33, 42]. While

these methods provide *valid* and *proximal* explanations for a model outcome, they fail to provide a *diverse* set of *sparse, non-trivial* explanations. Mothilal et al. [30] addressed the diversity problem by introducing a diversity constraint between randomly initialized counterfactuals (DICE). However, DICE shares the same problems as [3, 6] since perturbations are directly performed on the observed feature space and it is not designed to generate *non-trivial* explanations.

In this work we note that existing counterfactual generation methods tend to produce explanations that exaggerate or reduce the main attribute being classified, a property we call trivial explanation, and propose DiVE, a counterfactual explanation method that focuses on generating *non-trivial* explanations, which change the outcome of a classifier by modifying other attributes in the images, revealing spurious correlations or biases of the classifiers to ML practitioners. We provide a more exhaustive review of the related work in the Supplementary Material.

3. Proposed Method

We propose DiVE, an explainability method that can interpret an ML model by identifying the latent attributes that have the most effect on its output. Summarized in Figure 1a, DiVE uses an encoder, a decoder, and a fixed-weight ML model for which we have access to its gradients. In this work, we focus on a binary image classifier in order to produce visual explanations. DiVE consists of two main steps. First, the encoder and the decoder are trained in an unsupervised manner to approximate the data distribution on which the ML model was trained. Unlike PE [42], our encoder-decoder model does not need to train on the same dataset that the ML model was trained on. Second, we optimize a set of vectors ε_i to perturb the latent representation $\mathbf{z}$ generated by the trained encoder. The details of the optimization procedure are provided in the Supplementary Material. We use the following 3 main losses for this optimization: a counterfactual loss $\mathcal{L}_{\text{CF}}$ that attempts to fool the ML model, a proximity loss $\mathcal{L}_{\text{prox}}$ that constrains the explanations with respect to the number of changing attributes, and a diversity loss $\mathcal{L}_{\text{div}}$ that enforces the explainer to generate diverse explanations with only one confounding factor for each of them. Finally, we propose several strategies to mask subsets of dimensions in the latent space to prevent the explainer from producing trivial explanations. Next we explain the methodology in more detail.

3.1. Obtaining a counterfactual representation.

Given a data sample $\mathbf{x} \in \mathcal{X}$, its corresponding target $y \in \{0, 1\}$, and a potentially biased ML model $f(\mathbf{x})$ that approximates $p(y|\mathbf{x})$, our method finds a perturbed version of the same input $\tilde{\mathbf{x}}$ that produces a desired probabilistic outcome $\tilde{y} \in [0, 1]$, so that $f(\tilde{\mathbf{x}}) = \tilde{y}$. In order to produce semantically meaningful counterfactual explanations, we seek to learn a counterfactual model of the image generator with the corresponding latent representation $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^d$ of the input $\mathbf{x}$. Ideally, each dimension in $\mathcal{Z}$ represents a different semantic concept of the data, *i.e.*, the different dimensions are *disentangled*.

We note that performing counterfactual transformation on images is an unsolved problems with many challenges. Despite this, we move forward with a practical approach and verify empirically that the result is reasonable. Our general approach is inspired from Pawlowski et al. [34]. They show that when the causal graph is specified, it is possible to use a VAE to approximate counterfactual inference. In our case, we make the assumption that the underlying causal graph is a factorial z causing the image x. However, z is unobserved and cannot be identified in the general case [27]. Hence, we rely on β-TCVAE [4] with inductive bias to estimate a disentangle representation, which was shown to obtain competitive disentanglement in practice [27]. It follows the same encoder-decoder structure as the VAE [20], *i.e.*, the input data is first encoded by a neural network $q_\phi(z|\mathbf{x})$ parameterized by ϕ. Then, the input data is recovered by a decoder neural network $p_\theta(\mathbf{x}|z)$, parameterized by θ.

In addition to the β-TCVAE loss, we use the perceptual reconstruction loss from Hou et al. [15]. This replaces the pixel-wise reconstruction loss by a perceptual reconstruction loss, using the hidden representation of a pre-trained neural network R. Specifically, we learn a decoder D_θ generating an image, *i.e.*, $\tilde{\mathbf{x}} = D_\theta(\mathbf{z})$, and this image is re-encoded in a hidden representation: $\boldsymbol{h} = R(\tilde{\mathbf{x}})$, and compared to the original image in the same space using a normal distribution. Once trained, the weights of the encoder-decoder are fixed for the rest of the steps of our algorithm.

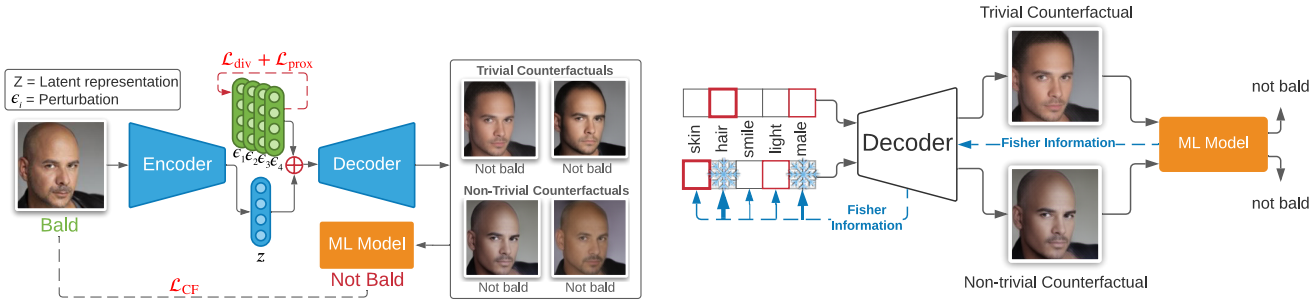
3.2. Interpreting the ML model

In order to find weaknesses in the ML model, DiVE searches for a collection of n latent perturbations $\{\boldsymbol{\epsilon}_i\}_{i=1}^n$ such that the decoded output $\tilde{\mathbf{x}}_i = D_\theta(\mathbf{z} + \boldsymbol{\epsilon}_i)$ yields a specific response from the ML model, *i.e.*, $f(\tilde{\mathbf{x}}) = \tilde{y}$ for any chosen $\tilde{y} \in [0, 1]$. We optimize $\boldsymbol{\epsilon}_i$'s by minimizing:

$$\begin{aligned}\mathcal{L}_{\text{DiVE}}(\mathbf{x}, \tilde{y}, \{\boldsymbol{\epsilon}_i\}_{i=1}^n) = &\textstyle\sum_i \mathcal{L}_{\text{CF}}(\mathbf{x}, \tilde{y}, \boldsymbol{\epsilon_i}) \\ &+ \lambda \cdot \textstyle\sum_i \mathcal{L}_{\text{prox}}(\mathbf{x}, \boldsymbol{\epsilon_i}) \\ &+ \alpha \cdot \mathcal{L}_{\text{div}}(\{\boldsymbol{\epsilon}_i\}_{i=1}^n), \end{aligned} \tag{1}$$

where λ, and α determine the relative importance of the losses. Minimization is performed with gradient descent and the complete algorithm can be found in the Supplementary Material. We now describe the different loss terms.

Counterfactual loss. The goal of this loss function is to identify a change of latent attributes that will cause the ML model f to change it's prediction. For example, in face recognition, if the classifier detects that there is a smile present whenever the hair is brown, then this loss function



(a) DiVE diagram

(b) Effect of the Fisher Information

Figure 1: **Left:** DiVE encodes the input image to explain into a latent representation z. Then z is perturbed by ϵ and decoded as counterfactual examples. During training, $\mathcal{L}_{\text{CF}}$ finds the set of ϵ that change the ML model classifier outcome while $\mathcal{L}_{\text{div}}$ and $\mathcal{L}_{\text{prox}}$ enforce that the samples are *diverse* and *proximal*. These are four valid counterfactuals from the experiment in Section 4.1. However, only the bottom row contains counterfactuals where the man is still bald as indicated by the oracle or a human. These counterfactuals identify a weakness in the ML model. **Right:** Fisher Information indicates the most important latent directions for the ML model, where importance is represented by the thickness of the blue line (hair in this example). We keep those directions fixed since they are usually trivial and thus explanations modify other attributes (red boxes).

is likely to change the hair color attribute. This is achieved by sampling from the decoder $\tilde{\mathbf{x}} = D_\theta(\mathbf{z}+\boldsymbol{\epsilon})$, and optimizing the binary cross-entropy between the target $\tilde{y}$ and the prediction $f(\tilde{\mathbf{x}})$:

$$\mathcal{L}_{\text{CF}}(\mathbf{x}, \tilde{y}, \boldsymbol{\epsilon}) = \tilde{y} \cdot \log(f(\tilde{\mathbf{x}})) + (1-\tilde{y}) \cdot \log(1-f(\tilde{\mathbf{x}})). \quad (2)$$

Proximity loss. The goal of this loss function is to constrain the reconstruction produced by the decoder to be similar in appearance and attributes as the input. It consists of the following two terms,

$$\mathcal{L}_{\text{prox}}(\mathbf{x}, \boldsymbol{\epsilon}) = ||\mathbf{x} - \tilde{\mathbf{x}}||_1 + \gamma \cdot ||\boldsymbol{\epsilon}||_1, \quad (3)$$

where γ is a scalar weighting the relative importance of the two terms. The first term ensures that the explanations can be related to the input by constraining the input and the output to be similar. The second term aims to identify a sparse perturbation to the latent space $\mathcal{Z}$ that confounds the ML model. This constrains the explainer to identify the least amount of attributes that affect the classifier's decision in order to produce *sparse* explanations.

Diversity loss. This loss prevents the multiple explanations of the model from being identical. For instance, if gender and hair color are spuriously correlated with smile, the model should provide images either with different gender or different hair color. To achieve this, we jointly optimize for a collection of n perturbations $\{\boldsymbol{\epsilon}_i\}_{i=1}^n$ and minimize their pairwise similarity:

$$\mathcal{L}_{\text{div}}(\{\boldsymbol{\epsilon}_i\}_{i=1}^n) = \sqrt{\sum_{i \neq j} \left(\frac{\boldsymbol{\epsilon}_i^T}{\|\boldsymbol{\epsilon}_i\|_2} \frac{\boldsymbol{\epsilon}_j}{\|\boldsymbol{\epsilon}_j\|_2} \right)^2}. \quad (4)$$

The method resulting of optimizing Eq. 1 (DiVE) results in *diverse* counterfactuals that are more *valid*, *proximal*, and *sparse*. However, it may still produce *trivial* explanations, such as exaggerating a smile to explain a smile classifier without considering other valuable biases in the ML model such as hair color. While the diversity loss encourages the orthogonality of the explanations, there might still be several latent variables required to represent all variations of smile.

Beyond trivial counterfactual explanations. To find *non-trivial* explanations, we propose to prevent DiVE from perturbing the most influential latent factors of $\mathcal{Z}$ on the ML model. We estimate the influence of each of the latent factors with the average Fisher information matrix:

$$\boldsymbol{F} = \mathbb{E}_{p(i)} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \mathbb{E}_{p(y|\mathbf{z})} \nabla_{\mathbf{z}} \ln p(y|\mathbf{z}) \; \nabla_{\mathbf{z}} \ln p(y|\mathbf{z})^T, \quad (5)$$

where $p(y=1|\mathbf{z}) = f(D_\theta(\mathbf{z}))$, and $p(y=0|\mathbf{z}) = 1-f(D_\theta(\mathbf{z}))$. The diagonal values of $\mathbf{F}$ express the relative influence of each of the latent dimensions on the classifier output. Since the most influential dimensions are likely to be related to the main attribute used by the classifier, we propose to prevent Eq. 1 from perturbing them in order to find more surprising explanations. Thus when producing n explanations, we sort $\mathcal{Z}$ by the magnitude of the diagonal, we partition it into n contiguous chunks that will be optimized for each of the explanations. We call this method $\text{DiVE}_{\text{Fisher}}$.

However, DiVE_{Fisher} does not guarantee that the different partitions of $\mathcal{Z}$ all the factors concerning a *trivial* attribute are grouped together. Thus, we propose to partition $\mathcal{Z}$ into subsets of latent factors that interact with each other when changing the predictions of the ML model. Such

interaction can be estimated using $\boldsymbol{F}$ as an affinity measure. We use spectral clustering [44] to obtain a partition of $\mathcal{Z}$. This partition is represented as a collection of masks $\{\boldsymbol{m}_i\}_{i=1}^{n}$, where $\boldsymbol{m}_i \in \{0,1\}^d$ represents which dimensions of $\mathcal{Z}$ are part of cluster i. Finally, these masks are used in Equation 1 to bound each $\boldsymbol{\epsilon}_i$ to its subspace, *i.e.*, $\boldsymbol{\epsilon}'_i = \boldsymbol{\epsilon}_i \circ \boldsymbol{m}_i$, where $\circ$ represents element wise multiplication. Since these masks are orthogonal, this effectively replaces $\mathcal{L}_{\text{div}}$. In Section 4, we highlight the benefits of this clustering approach by comparing to other baselines. We call this method $\text{DiVE}_{\text{FisherSpectral}}$.

4. Experimental Results

In this section, we first evaluate the described methods on their ability to identify diverse *non-trivial* explanations for image misclassifications made by the ML model (Section 4.1) and the out-of-distribution performance of DiVE (Section 4.1). In the following sections we validate the correctness of DiVE by evaluating its performance on 4 different aspects: (1) the *validity* of the generated explanations as well as the ability to discover biases within the ML model and the data (Section 4.2); (2) their *proximity* in terms of FID, latent space closeness, and face verification accuracy (Section 4.3); and (3) the *sparsity* of the generated counterfactuals (Section 4.4).

Experimental Setup. To align with [7, 18, 42], we perform experiments on the CelebA database [26]. CelebA is a large-scale dataset containing more than 200K celebrity facial images. Each image is annotated with 40 binary attributes such as "Smiling", "Male", and "Eyeglasses". These attributes allow us to evaluate counterfactual explanations by determining whether they could highlight spurious correlations between multiple attributes such as "lipstick" and "smile". In this setup, explainability methods are trained in the training set and ML models are explained on the validation set. The hyperparameters of the explainer are searched by cross-validation on the training set. We compare our method with xGEM [18] and PE [42] as representatives of methods that use an unconditional generative model and a conditional GAN respectively. We use the same train and validation splits as PE [42]. DiVE and xGEM do not have access to the labeled attributes during training.

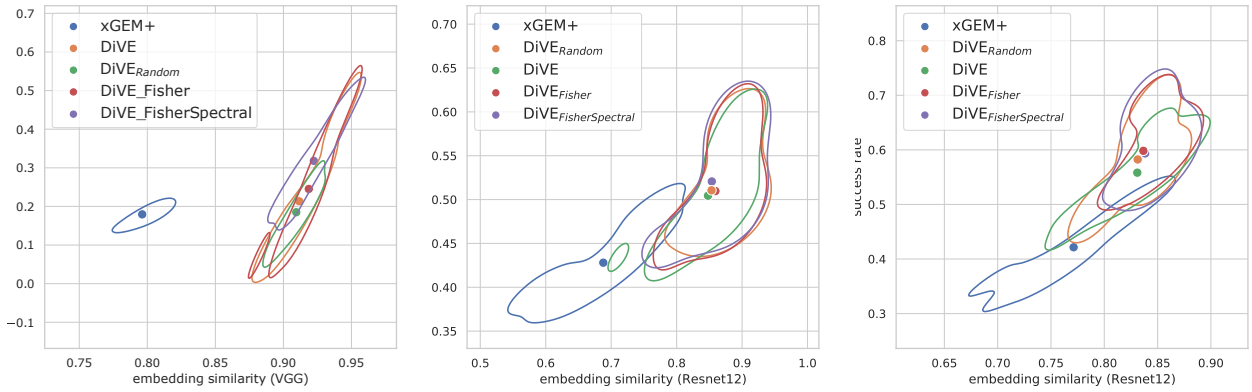
We test the out-of-distribution (OOD) performance of DiVE with the Synbols dataset [22]. Synbols is an image generator with characters from the Unicode standard and the wide range of artistic fonts provided by the open font community. This grants us better control the features present in each set when compared to CelebA. We generate 100K black and white of 32×32 images from 48 characters in the latin alphabet and more than 1K fonts. We use the character type to create disjoint sets for OOD training and we use the fonts to introduce biases in the data. We provide a sample of the dataset in the Supplementary Material.

We compare three version of our method and two ablated version to three existing methods. DiVE, resulting of optimizing Eq. 1. $\text{DiVE}_{\text{Fisher}}$, which extends DiVE by using the Fisher information matrix introduced in Eq. 5. $\text{DiVE}_{\text{FisherSpectral}}$, which extends $\text{DiVE}_{\text{Fisher}}$ with spectral clustering. We introduce two additional ablations of our method, DiVE−− and $\text{DiVE}_{\text{Random}}$. DiVE−− is equivalent to DiVE but using a pixel-based reconstruction loss instead of the perceptual loss. $\text{DiVE}_{\text{Random}}$ uses random masks instead of using the Fisher information. Finally, we compare our baselines with xGEM as described in Joshi et al. [18], xGEM+, which is the same as xGem but uses the same auto-encoding architecture as DiVE, and PE as described by Singla et al. [42]. For our methods, we provide implementation details, architecture description, and algorithm in the Supplementary Material.

4.1. Beyond trivial explanations

Previous works on counterfactual generations tend to produce *trivial* input perturbations to change the output of the ML model. That is, they tend to increase/decrease the presence of the attribute that is intended to be classified. For instance, in Figure 3 all the explainers put a smile on the input face in order to increase the probability for "smile". While that is correct, this explanation does not provide much insight about the potential weaknesses of the ML model. Instead, in this work we emphasize producing non-trivial explanations that are different from the main attribute that the ML model has been trained to identify. These kind of explanations provide more insight about the factors that affect the classifier and thus provide cues on how to improve the model or how to fix incorrect predictions.

To evaluate this, we propose a new benchmark that measures a method's ability to generate *valuable* explanations. For an explanation to be valuable, it should 1) be misclassified by the ML model (*valid*), 2) not modify attributes intended to be classified by the ML model (*non-trivial*), and 3) not have diverged too much from the original sample (*proximal*). A misclassification provides insights into the weaknesses of the model. However, the counterfactual is even more insightful when it stays close to the original image as it singles-out spurious correlations learned by the ML model. Because it is costly to provide human evaluation of an automatic benchmark, we approximate both the proximity and the real class with the VGGFace2-based oracle. We choose the VGGFace2 model as it is less likely to share the same biases as the ML model, since it was trained for a different task than the ML model with an order of magnitude more data. We conduct a human evaluation experiment in the Supplementary Material, and we find a significant correlation between the oracle and the human predictions. For 1) and 2) we deem that an explanation is successful if the ML



(a) CelebA (ID) (b) Synbols (ID) (c) Synbols (OOD)

Figure 2: **Beyond trivial explanations.** Rate of successful explanations (y-axis) against embedding similarity (x-axis) for all methods. The most valuable explanations are in the top-right corner. We ran an hyperparameter sweep and denote the mean of the performances with a dot. The curves are computed with KDE. The left plot shows the performance on CelebA and the other two plots shows the performance for in-distribution (ID) and out-of-distribution (OOD) experiments on Synbols. All DiVE methods outperform xGEM+ on both metrics simultaneously when conditioning on *successful counterfactuals*. In both experiments, $\text{DiVE}_{\text{Fisher}}$ and $\text{DiVE}_{\text{FisherSpectral}}$ improve the performance over both $\text{DiVE}_{\text{Random}}$ and DiVE.

model and the oracle make different predictions about the counterfactual. *E.g.*, the top counterfactuals in Figure 1a are not deemed successful explanations because both the ML model and the oracle agree on its class, however the two in the bottom row are successful because only the oracle made the correct prediction. These explanations where generated by $\text{DiVE}_{\text{FisherSpectral}}$. As for 3) we measure the proximity with the cosine distance between the sample and the counterfactual in the feature space of the oracle.

We test all methods from Section 4 on a subset of the CelebA validation set described in the Supplementary Material. We report the results of the full hyperparameter search. The vertical axis shows the success rate of the explainers, *i.e.*, the ratio of valid explanations that are non-trivial. This is the misclassification rate of the ML model on the explanations. The dots denote the mean performances and the curves are computed with Kernel Density Estimation (KDE). On average, DiVE improves the similarity metric over xGEM+ highlighting the importance of disentangled representations for identity preservation. Moreover, using information from the diagonal of the Fisher Information Matrix as described in Eq. 5 further improves the explanations as shown by the higher success rate of $\text{DiVE}_{\text{Fisher}}$ over DiVE and $\text{DiVE}_{\text{Random}}$. Thus, preventing the model from perturbing the most influential latent factors helps to uncover spurious correlations that affect the ML model. Finally, the proposed spectral clustering of the full Fisher Matrix attains the best performance validating that the latent space partition can guide the gradient-based search towards better explanations. We reach the same conclusions in Table 3, where we provide a comparison with PE for the attribute "Young". In addition, we provide results for a version of xGEM+ with more disentangled latent factors (xGEM++). We find that disentangled representations provide the explainer with a more precise control on the semantic concepts being perturbed, which increases the success rate of the explainer by 16%.

Out-of-distribution generalization. In the previous experiments, the generative model of DiVE was trained on the same data distribution (*i.e.*, CelebA faces) as the ML model. We test the out-of-distribution performance of DiVE by training its auto-encoder on a subset of the latin alphabet of the Synbols dataset [22]. Then, counterfactual explanations are produced for a different disjoint subset of the alphabet. To evaluate the effectiveness of DiVE in finding biases on the ML model, we introduce spurious correlations in the data. Concretely, we assign a different set of fonts to each of the letters in the alphabet as detailed in the Supplementary Material. In-distribution (ID) results are reported in Figure 2b for reference, and OD results are reported in Figure 2c. We observe that DiVE is able to find valuable counterfactuals even when the VAE was not trained on the same data distribution. Moreover, results are consistent with the CelebA experiment, with DiVE outperforming xGEM+ and Fisher information-based methods outperforming the rest.

4.2. Validity and bias detection

We evaluate DiVE's ability to detect biases in the data. We follow the same procedure as PE [42] and train two binary classifiers for the attribute "Smiling". The first one is trained on a biased version of CelebA where males are smil-

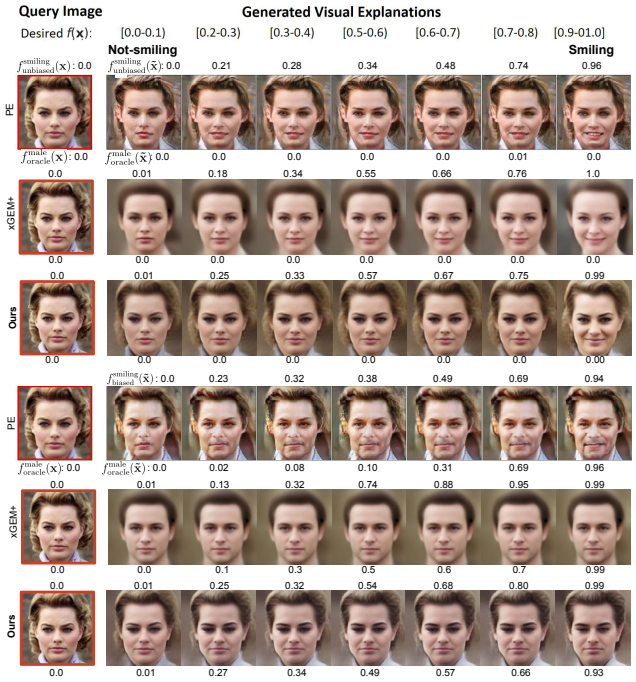


Figure 3: **Bias detection experiment.** Columns present explanations for a target "Smiling" probability interval. Rows contain explanations produced by PE [42], xGEM+ and our DiVE. (a) of a gender-unbiased classifier, and (b) a gender-biased "Smile" classifier. The classifier output probability is displayed on top of the images while the oracle prediction for gender is displayed at the bottom.

ing and females are not smiling (f_{biased}). This reflects an existing bias in the data gathering process where female are usually expected to smile [7, 13]. The second one is trained on the unbiased version of the data ($f_{unbiased}$). Both classifiers are evaluated on the CelebA validation set. Also following Singla et al. [42], we train an oracle classifier (f_{oracle}) based on VGGFace2 [2] which obtains perfect accuracy on the gender attribute. The hypothesis is that if "Smiling" and "Gender" are confounded by the classifier, so should be the explanations. Therefore, we could identify biases when the generated examples not only change the target attribute but also the confounded one. To generate counterfactuals, DiVE produces perturbations until it changes the original prediction of the classifier ("Smiling" to "Non-Smiling"). As described by Singla et al. [42] only *valid* explanations are considered, i.e. those that change the original prediction of the classifier.

We follow the procedure introduced in [18, 42] and report a confounding metric for bias detection in Table 1. The columns *Smiling* and *Non-Smiling* indicate the target class for counterfactual generation. The rows *Male* and *Female* contain the proportion of counterfactuals that are classified by the oracle as "Male" and "Female". We can see that the generated explanations for f_{biased} are classified more often as "Male" when the target attribute is "Smiling", and "Female" when the target attribute is "Non-Smiling". The confounding metric, denoted as *overall*, is the fraction of generated explanations for which the gender was changed with respect to the original image. It thus reflect the magnitude of the the bias as approximated by the explainers.

Singla et al. [42] consider that a model is better than another if the confounding metric is the highest on f_{biased} and the lowest on f_{unbiased}. However, they assume that f_{biased} always predicts the "Gender" based on "Smile". Instead, we propose to evaluate the confounding metric by comparing it to the empirical bias of the model, denoted as ground truth in Table 1. Details provided in the Supplementary Material.

We observe that DiVE is more successful than PE at detecting biases although the generative model of DiVE was not trained with the biased data. While xGEM+ has a higher success rate at detecting biases in some cases, it produces lower-quality images that are far from the input. In Figure 3, we provide samples generated by our method with the two classifiers and compare them to PE and xGEM+. We found that gender changes with the "Smiling" attribute with f_{biased} while for f_{unbiased} it stayed the same. In addition, we also observed that for f_{biased} the correlation between "Smile" and "Gender" is higher than for PE. It can also be observed that xGEM+ fails to retain the identity of the person in $\mathbf{x}$ when compared to PE and our method. Qualitative results are reported in Figure 3.

4.3. Counterfactual Explanation Proximity

We evaluate the *proximity* of the counterfactual explanations using FID scores [14] on CelebA as described by Singla et al. [42] (we observed similar results on MNIST and CIFAR [21, 23]). The scores are based on the target attributes "Smiling" and "Young", and are divided into 3 categories: *Present*, *Absent*, and *Overall*. *Present* considers explanations for which the ML model outputs a probability greater than 0.9 for the target attribute. *Absent* refers to explanations with a probability lower than 0.1. *Overall* considers all the successful counterfactuals, which changed the original prediction of the ML model.

We report these scores in Table 2 for all 3 categories. DiVE produces the best quality counterfactuals, surpassing PE by 6.3 FID points for the "Smiling" target and 19.6 FID points for the "Young" target in the *Overall* category. DiVE obtains lower FID than xGEM+ which shows that the improvement not only comes from the superior architecture of our method. Further, there are two other factors that explain the improvement of DiVE's FID. First, the β-TCVAE decomposition of the KL divergence improves the disentanglement ability of the model while suffering less reconstruction degradation than the VAE. Second, the perceptual loss makes the image quality constructed by DiVE to be comparable with that of the GAN used in PE. Additional exper-

Table 1: **Bias detection experiment.** Ratio of generated counterfactuals classified as "Smiling" and "Non-Smiling" for a classifier biased on gender (f_{biased}) and an unbiased classifier (f_{unbiased}). Bold indicates *Overall* closest to *Ground truth* (detailed in the Appendix).

ML model		Target label: Smiling PE	Smiling xGEM+	Smiling DiVE	Non-Smiling PE	Non-Smiling xGEM+	Non-Smiling DiVE
f_{biased}	Male	0.52	0.94	0.89	0.18	0.24	0.16
	Female	0.48	0.06	0.11	0.82	0.77	0.84
	Overall	0.12	**0.29**	0.22	0.35	0.33	**0.36**
	Ground truth		0.75			0.67	
f_{unbiased}	Male	0.48	0.41	0.42	0.47	0.38	0.44
	Female	0.52	0.59	0.58	0.53	0.62	0.57
	Overall	**0.07**	0.13	0.10	0.08	0.15	**0.07**
	Ground truth		0.04			0.00	

Table 2: FID of DiVE compared to xGEM [18], Progressive Exaggeration (PE) [42], xGEM trained with our backbone (xGEM+), and DiVE trained without the perceptual loss (DiVE−−)

Target Attribute	xGEM	PE	xGEM+	DiVE−−	DiVE
	Smiling				
Present	111.0	46.9	67.2	54.9	**30.6**
Absent	112.9	56.3	77.8	62.3	**33.6**
Overall	106.3	35.8	66.9	55.9	**29.4**
	Young				
Present	115.2	67.6	68.3	57.2	**31.8**
Absent	170.3	74.4	76.1	51.1	**45.7**
Overall	117.9	53.4	59.5	47.7	**33.8**

iments in the Supplementary Material show that DiVE is more successful at preserving the identity of the faces than PE and xGEM and thus at producing feasible explanations. These results suggest that the combination of disentangled latent features and the regularization of the latent features help DiVE to produce the minimal perturbations of the input that produce a successful counterfactual.

In Figure 3 we show qualitative results obtained by targeting different probability ranges for the output of the ML model as described in PE. DiVE produces more natural-looking facial expressions than xGEM+ and PE. Additional results for "Smiling" and "Young" are provided in the Supplementary Material.

4.4. Counterfactual Explanation Sparsity

We quantitatively compare the amount of valid and sparse counterfactuals provided by different baselines. Table 3 shows the results for a classifier model trained on the attribute "Young" of the CelebA dataset. The first row shows the number of attributes that each method change in average to generate a valid counterfactual. Attribute changes is measured from the output of the with the VGGFace2-based oracle. Methods that require to change less attributes are likely to be actionable by a user. We observe that DiVE changes less attributes on average than xGEM+. $\text{DiVE}_{\text{FisherSpectral}}$ is the method that changes less attributes. To better understand the effect of disentangled representations, we also report results for a version of xGEM+ with the β-TCVAE backbone (xGEM++). We do not observe significant effects on the sparsity of the counterfactuals. In fact, a fine-grained decomposition of concepts in the latent space could lead to lower the sparsity.

Table 3: Average number of attributes changed per explanation and percentage of non-trivial explanations. This experiment evaluates the counterfactuals generated by different methods for an ML model trained on the attribute "Young" of the CelebA dataset. xGEM++ is xGEM+ using β-TCVAE as generator.

	PE [42]	xGEM+ [18]	xGEM++	DiVE	$\text{DiVE}_{\text{Fisher}}$	$\text{DiVE}_{\text{FisherSpectral}}$
Attr. change	03.74	06.92	06.70	04.81	04.82	04.58
Non-trivial (%)	05.12	18.56	34.62	43.51	42.99	51.07

5. Limitations and Future Work

This work shows that a good generative model can provide interesting insights on the biases of an ML model. However, this relies on a properly disentangled representation. In the case where the generative model is heavily entangled it would fail to produce explanations with a sparse amount of features. However, our approach can still tolerate a small amount of entanglement, yielding a small decrease in interpretability. We expect that progress in identifiability [19, 28] will increase the quality of representations. With a perfectly disentangled model, our approach could still miss some explanations or biases. *E.g.*, with the spectral clustering of the Fisher, we group latent variables and only produce a single explanation per group in order to present explanations that are conceptually different. This may leave behind some important explanations, but the user can simply increase the number of clusters or the number of explanations per clusters for a more in-depth analysis.

In addition, finding the optimal hyperparameters for the VAE and their OOD generalization is an open problem. If the generative model is trained on biased data, one could expect the counterfactuals to be biased as well. However, as shown in Figure 2c, our model still finds non-trivial explanations when applied OOD.

Although the generative model plays an important role to produce valuable counterfactuals in the image domain, our work could be extended to other domains. For example, Eq. 1 could be applied on tabular data by directly optimizing observed features instead of latent factors of a VAE. However, further work would be needed to adapt DiVE to produce perturbations on discrete and categorical variables.

References

[1] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 2018. 2

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. 7

[3] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2019. 1, 2, 3

[4] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018. 2, 3

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 2

[6] P. Dabkowski and Y. Gal. Real time image saliency for black box classifiers. *arXiv preprint arXiv:1705.07857*, 2017. 2, 3

[7] E. Denton, B. Hutchinson, M. Mitchell, and T. Gebru. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*, 2019. 2, 5, 7

[8] R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *International Conference on Computer Vision*, 2017. 1, 2

[9] Y. Gal, J. Hron, and A. Kendall. Concrete dropout. In *Advances in neural information processing systems*, 2017. 1

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 2

[11] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual visual explanations. *arXiv preprint arXiv:1904.07451*, 2019. 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 2016. 2

[13] D. Hestenes and I. Halloun. Interpreting the force concept inventory: A response to march 1995 critique by huffman and heller. *The physics teacher*, 33(8):502–502, 1995. 7

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 2, 7

[15] X. Hou, L. Shen, K. Sun, and G. Qiu. Deep feature consistent variational autoencoder. In *Winter Conference on Applications of Computer Vision*, 2017. 3

[16] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Computer Vision and Pattern Recognition Workshops*, 2017. 2

[17] F. V. Jensen et al. *An introduction to Bayesian networks*, volume 210. UCL press London, 1996. 2

[18] S. Joshi, O. Koyejo, B. Kim, and J. Ghosh. xgems: Generating examplars to explain black-box models. *arXiv preprint arXiv:1806.08867*, 2018. 1, 2, 5, 7, 8

[19] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020. 8

[20] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3

[21] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7

[22] A. Lacoste, P. Rodríguez López, F. Branchaud-Charron, P. Atighehchian, M. Caccia, I. H. Laradji, A. Drouin, M. Craddock, L. Charlin, and D. Vázquez. Synbols: Probing learning algorithms with synthetic datasets. *Advances in Neural Information Processing Systems*, 33, 2020. 5, 6

[23] Y. LECUN. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*. URL https://ci.nii.ac.jp/naid/10027939599/en/. 7

[24] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 2

[25] S. Liu, B. Kailkhura, D. Loveland, and Y. Han. Generative counterfactual introspection for explainable deep learning. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE, 2019. 2

[26] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, 2015. 2, 5

[27] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019. 3

[28] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020. 8

[29] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 2017. 1, 2

[30] R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Conference on Fairness, Accountability, and Transparency*, 2020. 1, 3

[31] J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972. 2

[32] N. Papernot and P. McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018. 2

[33] M. Pawelczyk, K. Broelemann, and G. Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pages 3126–3132, 2020. 2

[34] N. Pawlowski, D. C. Castro, and B. Glocker. Deep structural causal models for tractable counterfactual inference. *arXiv preprint arXiv:2006.06485*, 2020. 3

[35] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986. 2

[36] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*, 2016. 1, 2

[37] C. Russell. Efficient search for diverse coherent explanations. In *Conference on Fairness, Accountability, and Transparency*, 2019. 1

[38] A. Sauer and A. Geiger. Counterfactual generative networks. *arXiv preprint arXiv:2101.06046*, 2021. 2

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision*, 2017. 1, 2

[40] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*, 2017. 1, 2

[41] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[42] S. Singla, B. Pollack, J. Chen, and K. Batmanghelich. Explanation by progressive exaggeration. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 5, 6, 7, 8

[43] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 2

[44] X. Y. Stella and J. Shi. Multiclass spectral clustering. In *null*, page 313. IEEE, 2003. 5

[45] A. Van Looveren, J. Klaise, G. Vacanti, and O. Cobb. Conditional generative models for counterfactual explanations. *arXiv preprint arXiv:2101.10123*, 2021. 2

[46] F. Yang, N. Liu, M. Du, and X. Hu. Generative counterfactuals for neural networks via attribute-informed perturbation. *arXiv preprint arXiv:2101.06930*, 2021. 2

[47] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 2

[48] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition*, 2016. 2