

# Locally Explaining Prediction Behavior via Gradual Interventions and Measuring Property Gradients

Niklas Penzel<sup>1</sup> Joachim Denzler<sup>1</sup>

<sup>1</sup>Computer Vision Group, Friedrich Schiller University Jena, Germany

{niklas.penzel, joachim.denzler}@uni-jena.de

## Abstract

*Deep learning models achieve high predictive performance but lack intrinsic interpretability, hindering our understanding of the learned prediction behavior. Existing local explainability methods focus on associations, neglecting the causal drivers of model predictions. Other approaches adopt a causal perspective but primarily provide global, model-level explanations. However, for specific inputs, it's unclear whether globally identified factors apply locally. To address this limitation, we introduce a novel framework for local interventional explanations by leveraging recent advances in image-to-image editing models. Our approach performs gradual interventions on semantic properties to quantify the corresponding impact on a model's predictions using a novel score, the expected property gradient magnitude. We demonstrate the effectiveness of our approach through an extensive empirical evaluation on a wide range of architectures and tasks. First, we validate it in a synthetic scenario and demonstrate its ability to locally identify biases. Afterward, we apply our approach to investigate medical skin lesion classifiers, analyze network training dynamics, and study a pre-trained CLIP model with real-life interventional data. Our results highlight the potential of interventional explanations on the property level to reveal new insights into the behavior of deep models.<sup>1</sup>*

## 1. Introduction

Modern deep learning models are complex data-centric systems that achieve high predictive performance but lack intrinsic interpretability. Hence, many explainability (XAI) methods were proposed to interpret trained model behavior, especially for vision models. Post-hoc XAI includes local methods, often generating pixel-wise attributions, e.g., [39, 53, 63, 64, 66] and global methods focused on human interpretable concepts or properties [9, 11, 16, 30, 32, 40, 52, 58, 69]. Unfortunately, global insights can be deceiving

for individual inputs. On the local level, properties can be occluded or overshadowed by independent visual elements.

To address this limitation, we introduce a novel approach to locally explain neural network prediction behavior based on input interventions at the property level. We propose to leverage recent breakthroughs in image-to-image editing models, e.g., [5, 15, 41]. These models are trained conditionally using Classifier-Free Guidance (CFG) scaling [26]. Hence, we can gradually control the alignment with the corresponding interventional instruction during inference. Our key insight is that this paradigm facilitates the smooth transition between two property states (see Fig. 1, top row), also for complex features. Consequently, by utilizing CFG scaling, we intervene on semantic properties for an individual input and study the shift in prediction behavior of trained neural networks (see Fig. 1, bottom row). To quantify the impact of a property on a model, we propose approximating the corresponding expected property gradient magnitude. This score is naturally connected to the intuition of measuring the change along the property axis in Fig. 1. Additionally, our expected property gradient magnitude can be seen as an extension of the causal concept effect [18] for gradual interventions. Finally, to ensure robustness, we suggest a corresponding permutation significance test.

We empirically validate our framework for local interventional explanations on a wide variety of architectures and tasks. First, we explore a biased scenario (cats vs. dogs [10]) where we synthetically correlate a property (fur color) with the label to validate that our approach can locally identify the causal factor (see Fig. 1). Additionally, our gradual interventions allow for a direct quantification of how the outputs change in response to shifts in the selected property, a key distinction from local methods highlighting image regions, e.g., [13, 19, 31, 39, 48, 53, 63, 64, 66]. Further, our derived score quantitatively outperforms baselines for indicating locally biased behavior during interventions. We corroborate these findings for real-world skin lesion classification, analyzing a known bias. Afterward, we study the training dynamics of eight modern classifiers by tracking a property correlated with the label. Here we find oscillations

<sup>1</sup>Project page: <https://proppgrad.github.io/>

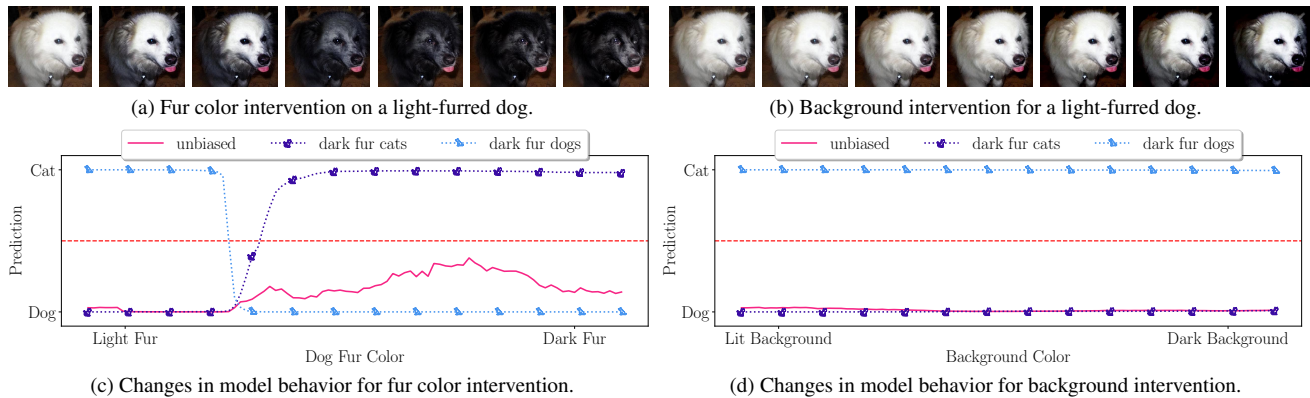


Figure 1. Fig. 1a and Fig. 1b demonstrate our approach to study model predictions through gradual interventions here for two properties: fur color and background illumination in a cat vs. dog classification task. Using [15], we synthetically adjust these properties, i.e., darkening the fur or the background, while tracking prediction changes. Specifically, Fig. 1c and Fig. 1d show the resulting shifts in model outputs for three networks: an unbiased one, one trained only on dark-furred cats and light-furred dogs, and one trained on the opposite bias. The red line marks the decision threshold, where predictions flip between classes during the intervention.

also in later epochs, depending on the weight initialization. Finally, we demonstrate that our model-agnostic approach works with diverse sources of interventions by capturing real-life interventional data to analyze a CLIP [50] model.

Our key contributions can be summarized as follows: (1) We introduce a new framework for local network prediction explanations based on gradual interventions, e.g., by leveraging Classifier-Free Guidance (CFG) scaling [26]. (2) We derive a novel score to quantify the shift in model behavior with respect to a property by approximating the expected property gradient magnitude. (3) We provide a corresponding hypothesis test to verify statistical significance. (4) We conduct experiments on a wide range of architectures and tasks to demonstrate the effectiveness of our approach.

## 2. Related Work

Many methods to derive local explanations aim to find important regions in the input, e.g., [39, 53, 63, 64, 66, 70]. A subset of these methods, most closely related to our approach, employs input perturbations to estimate importance. These perturbations or interventions are often done by replacing patches and, therefore, occluding parts of the input. Such occlusion patches can be constructed using noise, e.g., [47, 70]. Other approaches use similar image regions [71] or generative infilling [8, 31, 33]. Similarly related are methods that use causal terminology and generate visual counterfactual explanations [65] by posing questions of the form “Which parts of the input need to change to result in a specific prediction?” Examples of this approach include [2, 13, 19, 31, 48]. Nevertheless, such visual explanations necessitate additional semantic interpretations by experts to identify specific properties responsible for the measured importance. In contrast, our approach explains the prediction

behavior directly on the level of semantic properties.

Related in that regard are, often global, explanation methods, e.g., [9, 11, 16, 23, 30, 32, 34, 40, 52, 58–60, 69], which discover abstract concepts learned by a trained model. However, these methods require direct access to the model parameters or probing datasets. Further, while they are often explorative, they have difficulty determining whether a certain property is unused and can suffer from confounding, e.g., [32], see [18]. The approach described in [52] can test for the usage of human-defined properties by trained models. To do this, they assume usage and either confirm or reject the null hypothesis using conditional independence tests. However, the results are binary and, unfortunately, do not allow actionable interpretations of the changes in prediction behavior on a local level. While other works, e.g., [7, 46], use probing datasets to tackle these limitations, the explanations are strictly associational. While we similarly test for significance, we propose an interpretable score for the local impact of specific properties. Additionally, our approach is inherently interventional.

Related to the interventional nature of our work is [6]. The authors generate synthetic data with selected interventions to investigate emotion classifiers. We extend their work to more general input properties and provide a structured way to generate local interventional explanations. Other synthetic analysis datasets, e.g., [3, 25, 51], can be used to study model behavior. In contrast, we directly intervene in inputs to remove the domain shift necessary for the synthetic analysis of pre-trained models. The related approach [30] explains models globally by intervening on image semantics, while we focus in our work on locally attributing the importance of properties. Regarding the estimation of the impact with respect to a property, our work is most closely related to [18]. In fact, our measure can be

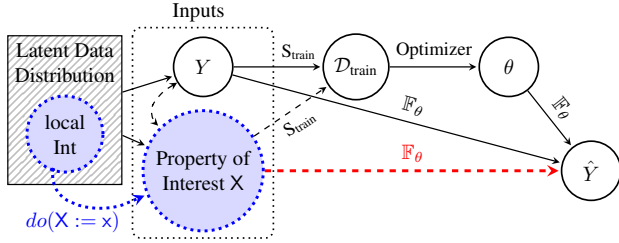


Figure 2. Our structural causal model [43] for the property dependence of trained networks. Dashed connections potentially exist depending on the specific task/property combination, and the sampled ( $S_{train}$ ) data. In this work, we study the **red dashed link** between  $X$  and  $\hat{Y}$ . By intervening on  $X$ , i.e.,  $do(X := x)$ , we induce changes in  $\hat{Y}$ , which are fully determined by the network  $\mathbb{F}_\theta$ .

seen as an extension of the causal concept effect for gradual interventions. Lastly, we discuss the interaction with Pearl’s causal hierarchy and the causal hierarchy theorem [4], specifically for visual interventions [42]. Other works study the link between causality and explanations by intervening on training hyperparameters [29] or by casting explanations as falsifiable hypotheses to be verified [61].

### 3. Method

#### 3.1. Causal Preliminaries

We employ structural causal models (SCMs) [43] to describe the data-generating process underlying our analysis. SCMs provide a flexible framework to model complex relationships between variables. In this work, we aim to understand the decision-making and prediction behavior of deep neural networks for individual inputs. Consequently, we model the outputs of trained networks as results in a data-generating system and visualize our proposed SCM in Fig. 2. In there, the network outputs  $\hat{Y}$  are deterministically produced by a parameterized model  $\mathbb{F}_\theta$ . The weights are optimized on a collection of training data  $\mathcal{D}_{train}$ . This training data is not arbitrary. Instead, the sampled inputs strongly depend on task-related reference annotations  $Y$ .

In the literature, different alternatives to describe the inputs in such a system are discussed, e.g., [6, 18, 29, 52]. In our work, we follow [6, 52] and model inputs as a possibly infinite collection of semantic properties. These properties are not necessarily independent and can be causally related or spuriously correlated. Note that the reference annotation  $Y$  is such a semantic property, and  $\mathbb{F}_\theta$  primarily aims to extract it. Nevertheless, to interpret the prediction behavior of  $\mathbb{F}_\theta$ , we study the influence of other properties of interest  $X$  on  $\hat{Y}$ . To do this, we propose to intervene in individual input properties and measure the changes induced in the outputs.

#### 3.2. Why Do We Need Interventions?

Causal insights can be hierarchically ordered in the so-called causal ladder [43]. This ladder, formally Pearl’s Causal Hierarchy (PCH) [4], contains three distinct levels: associational, interventional, and counterfactual (see [4] for a formal definition). The first level, associational, is characterized by correlations observed in a given system. It focuses on statistical patterns and relationships within the data. The second level, interventional, involves actively changing variables within the system to study the resulting effects. This is formally represented using the  $do$ -operator [43], which allows researchers to examine the causal impact of interventions. The third level, counterfactual, deals with hypothetical scenarios, where researchers consider the potential outcome if an intervention had been made, given specific observations. Crucially, the causal hierarchy theorem [4, Thm. 1] states that the three levels are distinct, and the PCH almost never collapses in the general case. Hence, to answer questions of a certain PCH level, data from the corresponding level is needed [4, Cor. 1].

Consequently, our work falls into the second level and generates insights beyond associations. Related works on the interventional level either focus on pixel attributions, e.g., [8, 47, 70, 71], or global explanations of semantic properties [18, 30]. Such globally identified causal factors do not necessarily hold locally. To be specific, for a particular input, selected properties could be occluded or overshadowed. Our approach closes this gap and provides local interventional insights for specific semantic properties.

#### 3.3. Generating Interventional Data

To locally explain prediction behavior in the vision domain, we propose intervening directly on a property of interest  $X$  ( $do(X := x)$  in Fig. 2). To perturb  $X$  in an image, we identify three options: Capturing new interventional data, designing synthetic interventions, and interventions via generative models. The first two approaches involve collecting new interventional data [4] and are suited for specialized tasks in complex domains. And while we empirically assess them in Sec. 4, we agree with [18] and argue that generative models offer broad applicability with reduced human labor.

Consequently, for local interventions, we propose to leverage recent breakthroughs in image-to-image editing models, e.g., [5, 15, 41]. These models are based on latent diffusion [55] and align inputs with text prompts through classifier-free guidance (CFG) scaling [26]. Following [5] for timesteps  $t$ , CFG scaling utilizes

$$\bar{e}(z_t, c_T) = e(z_t, \emptyset) + s_T(e(z_t, c_T) - e(z_t, \emptyset)), \quad (1)$$

during inference, to steer a parameterized score network  $e$  away from the unconditional distribution,  $e(z_t, \emptyset)$ , when predicting the noise in latents  $z_t$ . Hence, increasing the

CFG scale  $s_T$  increases alignment with the conditioning text instruction  $c_T$ . In practice, [5, 15] include a second conditioning term, which we discuss in Appx. A.2.

Crucially, in Eq. (1),  $e$  is optimized jointly as a conditional model [26]. Therefore, the generative model learns to interpolate between property states during training. This is particularly important for achieving non-linear and gradual transitions, allowing us to study more complex properties. Hence, these prompting capabilities facilitate targeted and controllable interventions via text instructions. Note that our focus on semantic properties separates our approach from existing works providing visual counterfactual explanations, e.g., [13, 19, 31, 48] or see [65].

In addition, intervening in the input has distinct advantages. First, the explanations are model-agnostic and do not depend on a specific architecture. Second, we do not need access to a model’s weights, only its outputs. Finally, users can visually inspect the interventions and potentially include prior knowledge. Consequently, it allows for manual validation of interventional data, which is related to the idea of care sets proposed as a relaxation of the causal hierarchy theorem in [42]. We discuss alternatives to input space interventions in Appx. A.3. Having established how to intervene, we now turn our attention to measuring the impact of these interventions on a model’s predictions. Specifically, given interventional data with respect to a property  $X$ , we aim to quantify the corresponding changes in  $\hat{Y}$ .

### 3.4. Measuring Systematic Change

We utilize CFG scaling (Eq. (1)) to gradually intervene in a property of an original input image to generate interventional data. Next, to measure the changes induced in the network outputs, we approximate the magnitude of the gradient with respect to the property  $X$ , i.e.,  $|\nabla_X \mathbb{F}_\theta(I_X)|$ . Here,  $I_x$  is an input with a specific realization  $X = x$ .

Gradients as a measure of change or impact with respect to  $X$  are related to the causal concept effect [18] and can be seen as an extension for gradual interventions. We provide a detailed discussion of this connection in Appx. A.4. Nevertheless, given our gradual approach, this extension is important as periodical or parabolic effects of properties are potentially possible. To illustrate this possibility, consider the following example. Imagine a young person with brown hair, and suppose we gradually intervene on their hair color, transitioning from brown to gray to white. At first, an age classifier may become increasingly uncertain as the hair color changes, predicting an older age as the hair becomes grayer. However, once the hair is completely white, the classifier may again correctly predict a younger age due to hair color trends, e.g., platinum blonde or white hair amongst young people. By approximating gradient magnitudes for gradual variations in the hair color property, we can capture this non-linear shift in behavior.

Specifically, we subsample interventions and create a discrete ordered list of interventional inputs  $I_x$ , e.g., by using [15] and linearly increasing the CFG scale. We then compute the output of the trained model for each interventional input sample. Finally, assuming a set  $\mathfrak{X}$  of equidistant property realizations  $x$  (\*) [6], we approximate the expected gradient magnitude with respect to  $X$  with

$$\begin{aligned} \mathbb{E}_X[|\nabla_X \mathbb{F}_\theta(I_X)|] &= \int |\nabla_x \mathbb{F}_\theta(I_x)| \cdot p(x) dx \\ &\stackrel{(*)}{=} \frac{1}{|\mathfrak{X}|} \sum_{x \in \mathfrak{X}} |\nabla_x \mathbb{F}_\theta(I_x)|, \end{aligned} \tag{2}$$

where  $p(x)$  is the probability density of realizations  $x$ .

We refer to scores measured using Eq. (2) as expected property gradient magnitudes, or  $\mathbb{E}[|\nabla_X|]$  as a short-form. To accurately approximate Eq. (2) given our discrete list of intervened input samples, we employ Fornberg’s finite differences [14], as provided in [20]. The expected property gradient magnitude is an interpretable score of the systematic change with respect to variations in  $X$ . To illustrate this, a score of  $\mathbb{E}[|\nabla_X|] = 0.01$  indicates an average deviation of one percent for each discrete step of  $X$ .

Nevertheless,  $\mathbb{E}[|\nabla_X|]$  has limitations, and a high score does not necessarily indicate significance. In fact, noise can lead to high gradient magnitudes even if no systematic behavior exists. Hence, we need to differentiate between systematic and random changes in the prediction behavior.

### 3.5. Testing for Statistical Significance

A high effect size, measured with statistics such as  $\mathbb{E}[|\nabla_X|]$  or Pearson’s correlation [44], does not imply significance. Hence, to determine significance, we follow [6] and employ shuffle hypothesis testing, e.g., [17]. This approach compares a test statistic from the original observations to  $K$  randomly shuffled versions. Here, the interventional values of  $X$  and the corresponding model outputs constitute the original correspondence. We use  $\mathbb{E}[|\nabla_X|]$  as our test statistic, which connects our measure of behavior changes to the hypothesis test. Permuting the observations destroys the systematic relationship between  $X$  and the model outputs and facilitates approximating the null hypothesis (pseudo-code in Appx. A.5). In our experiments, we use a significance level 0.01 and perform 10K permutations.

## 4. Experiments

We demonstrate the effectiveness of our local interventional approach in various experiments. First, we validate it in a synthetic biased scenario, where we correlate a property with the label. Second, we investigate a realistic skin lesion classification task regarding a known bias. Afterward, we study the training dynamics of eight modern image classification models and, finally, a large pre-trained CLIP [50] model using real-life interventional data.

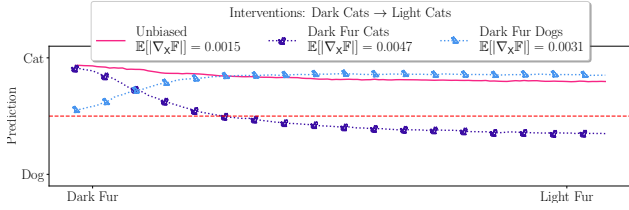


Figure 3. Average changes in model outputs (softmaxed cat-logit) for an intervention on the fur color in all **dark-furred cat** test images. Here, we use three models: an unbiased one, one trained on only dark-furred cats and light-furred dogs, and one trained on the opposite bias. The legend lists the mean  $\mathbb{E}[|\nabla_{\mathbf{x}}\mathbb{F}|]$  per model.

Table 1.  $\mathbb{E}[|\nabla_{\mathbf{x}}\mathbb{F}|]$  of the fur color and background property for our three CvD models. Additionally, we report significance ( $p < 0.01$ ) abbreviated as “S” and prediction flips denoted as “F”.

Model	Fur Color			Background		
	$\mathbb{E}[ \nabla_{\mathbf{x}}\mathbb{F} ]$	S	F	$\mathbb{E}[ \nabla_{\mathbf{x}}\mathbb{F} ]$	S	F
Unbiased	<b>.0099</b>	✓	✗	.00060	✓	✗
Dark Cats	<b>.0109</b>	✓	✓	.00006	✓	✗
Dark Dogs	<b>.0110</b>	✓	✓	.00013	✓	✗

#### 4.1. Synthetic Biased Scenario

We begin our empirical evaluation by constructing a biased scenario from the Cats vs. Dogs (CvD) dataset [10]. Using [35], we create two additional variations of the original distribution, where the fur color strongly correlates with the label. After manually verifying this approach, we obtain three training and test data splits: the original unbiased split, a split with only dark-furred dogs and light-furred cats, and the reverse. We hypothesize that models trained on the biased data splits will strongly rely on the fur color.

To test this hypothesis, we train a ConvMixer [68] model on each split until it achieves high performance (see Appx. B for concrete numbers). Biased models exhibit strong performance degradation on out-of-distribution test splits, while the unbiased model achieves consistent accuracy across all three scenarios. However, it is unclear whether the fur color is locally the dominant property driving predictions for individual inputs. Other properties, such as background illumination, may also influence model behavior. To investigate local prediction behavior, we use [15] to intervene on both fur color and background, increasing the CFG scale from 1 to 15. Finally, we quantify the impact using  $\mathbb{E}[|\nabla_{\mathbf{x}}\mathbb{F}|]$  and test for significance. We provide the full hyperparameters and additional ablations in Appx. B.

**Results:** Fig. 1 visualizes the model output behavior of all three classifiers for both the fur color intervention (Fig. 1a) and the background intervention (Fig. 1b) for an

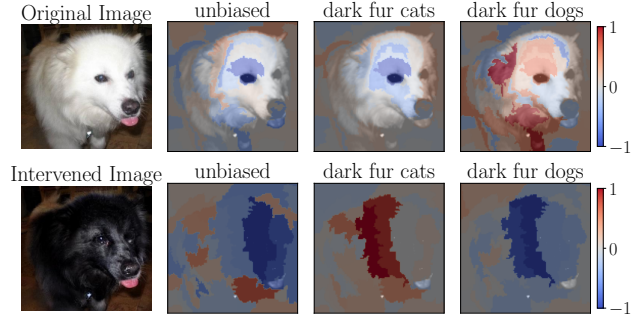


Figure 4. Local explanations using LIME [53] for the three CvD models (compare to Fig. 1). We explain with respect to the cat-logit, i.e., red areas indicate cat, while blue signals dog.

example. For the fur color intervention, we can confirm our hypothesis: both biased models flip their predictions to follow the change in fur color, locally disregarding the actual animal. While the unbiased model always predicts the correct class (dog), its logit for cat still increases. In contrast, the background intervention has a minimal effect, and no model crosses the prediction threshold.

To quantify these findings, we approximate the expected gradient magnitudes with respect to the intervened properties (see Tab. 1). For all models, we measure at least one order of magnitude higher  $\mathbb{E}[|\nabla_{\mathbf{x}}\mathbb{F}|]$  for the fur color compared to the background property, corroborating the observed behavior changes. Further, we measure the highest  $\mathbb{E}[|\nabla_{\mathbf{x}}\mathbb{F}|]$  of the fur color property for the two biased models, which coincides with the only observed prediction flips.

Next, to ensure robustness, we repeat the experiment with all test images again, intervening in the fur colors. Fig. 3 visualizes the average model outputs for the dark furred cat split, and we provide the full details, additional ablations for the remaining splits, background interventions, and a discussion of the intervention failure cases in Appx. B. Again, we observe the largest changes for the biased models, which align with flips in the prediction. Thus, our interventional approach successfully identifies the fur color as a local cause for the observed model outputs.

To stress its effectiveness, we compare our approach against local [39, 53, 63, 64, 66, 70] and global [52, 59, 69] XAI baselines (Appx. B.4, B.5). While global methods find influential properties, they do not quantify their local impact. Local attribution methods highlight important areas, but they require semantic interpretation. For example, although LIME explanations in Fig. 4 align with fur color for biased models, the distinction from the unbiased model is unclear. Interventions, i.e., the disparity between the top and bottom row in Fig. 4, can help interpret the results.

To formalize this comparison, we propose a quantitative task inspired by insertion/deletion tests [47]: predicting whether an intervention on a property will change the

Table 2. Mean accuracy ( $\uparrow$ ) and standard deviation in percent (%) of local XAI methods when predicting locally biased model behavior for an intervention. The first column denotes the dataset, i.e., Cats vs. Dogs [10] (CvD) and ISIC archive [1] (ISIC). For CvD, we evaluate the three ConvMixer [68] from the separate training datasets. We investigate the interventions discussed in Sec. 4.1 and Sec. 4.2, respectively.

	Model	Ours	G-CAM [63]	Int. Grad. [66]	Occlusion [70]	LIME [53]	K-SHAP [39]	DeepLift [64]
CvD	Unbiased	<b>86.13 <math>\pm</math> 2.0</b>	84.70 $\pm$ 2.1	84.77 $\pm$ 2.1	84.73 $\pm$ 2.0	84.70 $\pm$ 2.2	84.67 $\pm$ 2.1	84.83 $\pm$ 2.1
	Dark Cats Bias	<b>84.27 <math>\pm</math> 1.7</b>	61.43 $\pm$ 3.3	64.77 $\pm$ 1.8	67.70 $\pm$ 1.9	58.63 $\pm$ 1.9	60.87 $\pm$ 1.0	67.13 $\pm$ 2.2
	Dark Dogs Bias	<b>82.20 <math>\pm</math> 0.8</b>	64.20 $\pm$ 2.7	64.10 $\pm$ 2.2	67.10 $\pm$ 1.2	56.50 $\pm$ 2.1	58.50 $\pm$ 2.2	64.77 $\pm$ 2.1
ISIC	ResNet18 [21]	<b>95.50 <math>\pm</math> 1.2</b>	80.00 $\pm$ 4.0	78.63 $\pm$ 2.7	78.75 $\pm$ 4.0	76.12 $\pm$ 5.1	75.88 $\pm$ 4.6	76.88 $\pm$ 6.5
	EfficientNet-B0 [67]	<b>94.25 <math>\pm</math> 2.4</b>	79.00 $\pm$ 4.9	73.75 $\pm$ 5.6	75.37 $\pm$ 6.3	73.88 $\pm$ 5.1	72.75 $\pm$ 5.5	76.38 $\pm$ 6.5
	ConvNeXt-S [38]	<b>92.88 <math>\pm</math> 1.6</b>	78.00 $\pm$ 4.7	75.50 $\pm$ 3.2	74.88 $\pm$ 4.7	75.00 $\pm$ 3.2	74.75 $\pm$ 2.7	74.88 $\pm$ 5.8
	ViT-B/16 [12]	<b>95.12 <math>\pm</math> 1.6</b>	80.25 $\pm$ 3.4	75.87 $\pm$ 2.5	77.25 $\pm$ 3.5	76.00 $\pm$ 3.4	75.88 $\pm$ 2.9	77.50 $\pm$ 3.9

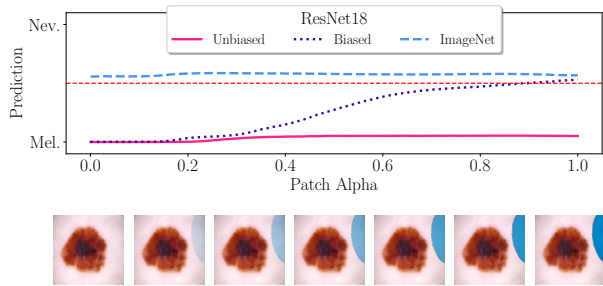


Figure 5. We visualize (top) how interventions targeting spurious colorful patches (bottom) [62] affect three ResNet18 models: trained on biased/unbiased skin lesion data and ImageNet weights.

model’s output. This task directly assesses if a method can indicate locally biased behavior. Further, this setup allows for a quantitative comparison to local baselines, given that our approach does not produce saliency maps but rather estimates the impact of a property directly using  $\mathbb{E}[|\nabla_{\mathcal{X}}|]$ . To adapt saliency methods, we measure the mean squared difference of the explanation pre- and post-intervention (e.g., between the two rows in Fig. 4). For a fair comparison, we select the optimal threshold maximizing the accuracy for both the local baselines and our score  $\mathbb{E}[|\nabla_{\mathcal{X}}|]$ . We repeat this task for the fur color property of 150 test images per class, resampling ten times to estimate standard deviations. Tab. 2 (top) summarizes the results and highlights the advantage of our approach. While all methods perform consistently well in predicting changes for the unbiased model,  $\mathbb{E}[|\nabla_{\mathcal{X}}|]$  clearly improves over baselines for models exhibiting local bias. Following this validation, we next analyze a real-world bias in skin lesion classification.

## 4.2. Skin Lesion Classification

In the domain of skin lesion classification (here, nevus/healthy vs. melanoma), a known bias is the correlation between colorful patches [62] and the nevus class. We assess how strongly this property is learned by four architectures: ResNet18 [21], EfficientNet-B0 [67], ConvNext-S

Table 3.  $\mathbb{E}[|\nabla_{\mathcal{X}}|]$  for colorful patch interventions [62] in skin lesion classifiers. We evaluate different models and training data.

Model	Training Data		
	Unbiased	Biased	ImageNet
ResNet18 [21]	.00061	<b>.00531</b>	.00062
EfficientNet-B0 [67]	.00018	<b>.00495</b>	.00066
ConvNeXt-S [38]	.00001	<b>.00519</b>	.00081
ViT-B/16 [12]	.00016	<b>.00208</b>	.00129

[38], and ViT-B/16 [12]. For each model, we start with ImageNet [57] weights. Then, we either fine-tune on biased skin lesion data (50% nevi with colorful patches [62]) or unbiased data (no patches) from the ISIC archive [1].

To demonstrate that our approach accommodates diverse sources of interventional data, we build on domain knowledge and intervene synthetically. Specifically, we blend segmented colorful patches [54] into melanoma images (see Fig. 5, bottom). We randomly sample ten correctly classified melanoma images and repeat interventions with five patches each. Appx. C contains detailed hyperparameters, predictive performances, and additional visualizations.

**Results:** The mean  $\mathbb{E}[|\nabla_{\mathcal{X}}|]$  in Tab. 3 show that models trained on biased data are most impacted by colorful patch interventions, indicating they learn the statistical correlation between the patches and the nevus class. Furthermore, the variants with ImageNet [57] weights show higher patch sensitivity than the unbiased skin lesion models. We hypothesize this is because learning color is beneficial for general-purpose pre-training, whereas the unbiased models learn to disregard patches and focus on the actual lesions.

These results are further corroborated in Fig. 5, where we visualize the average changes in model outputs for the ResNet18 [21] under the synthetic colorful patch intervention. Specifically, we observe that the biased model flips its prediction and incorrectly classifies the melanoma images as healthy. This highlights the ability of our interventional

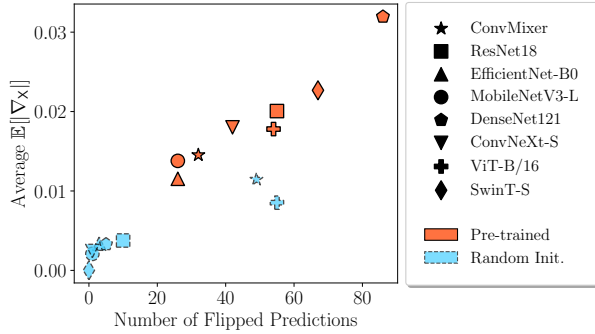


Figure 6. Average  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  visualized against the number of prediction flips for the training of various architectures with respect to the gray hair intervention. We display both ImageNet [57] pre-trained and randomly initialized models.

approach to investigate complex scenarios and use domain knowledge to provide actionable local explanations.

To further substantiate the benefit of our approach, we again compare against various local baselines in predicting whether an intervention (adding a colorful patch) will change the model’s output. We follow the setup described in Sec. 4.1, here using 100 melanoma test images, resampling ten times to estimate standard deviations. We report accuracies for the four architectures averaged over the bi-ased and unbiased training in Tab. 2 (bottom). Our score  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  significantly outperforms baseline methods for this task ( $p < 0.002$ ), highlighting its use case as a robust score to interpret local prediction behavior on the property level. However, our aim is not to replace saliency methods, but rather to offer a complementary, interventional viewpoint for analyzing local behavior. We demonstrate these capabilities in the following experiments.

### 4.3. Training Analysis

We investigate how the  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  of a property locally develops during the training of various architectures. This is an important question because it helps us understand how diverse models learn to represent and utilize properties in the data. Additionally, it is crucial to consider the impact of the initial parameters on the learned properties [45]. To address these questions, we select a range of convolutional and transformer-based architectures widely used in computer vision tasks [12, 21, 27, 28, 37, 38, 67, 68] (see Fig. 6). For all of these models, we train a randomly initialized and an ImageNet [57] pre-trained version for 100 epochs.

Regarding the corresponding task, we construct a binary classification problem from CelebA [36], following an idea proposed in [42]. Specifically, we utilize the attribute young as a label and split the data in a balanced manner. For this label, the gray hair property is negatively correlated [42],

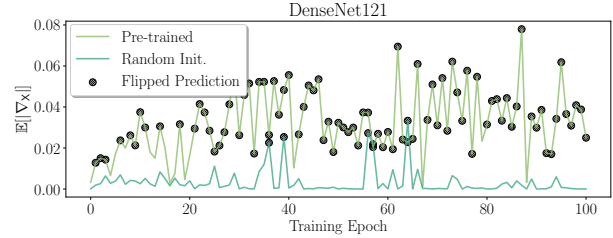


Figure 7.  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  development during training of DenseNet121 [28] models with respect to the gray hair property. We separate ImageNet pre-trained [57] and randomly initialized weights, and highlight epochs where predictions flip under the intervention.

and a well-performing classifier should learn this association during the training process. To study the training dynamics, we intervene on the hair color of a young test sample using [15] and calculate  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  after each epoch. We include detailed hyperparameters in Appx. D together with additional ablations and visualizations.

**Results:** In Fig. 6, we visualize the average  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  of the hair color for a local example over the training process for both pre-trained and randomly initialized models. Specifically, we display the average  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  against the observed flips in the prediction during the hair color intervention. In Appx. D.2, we include the concrete numbers (Tab. 12). Our analysis reveals two key insights:

First, for all architectures, the pre-trained variants locally exhibit higher  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  compared to the randomly initialized versions. This observation is consistent with the number of times the networks’ predictions flip during training. Notably, regarding prediction changes, the ConvMixer [68] and ViT [12] are outliers. However, these two models also demonstrate the highest measured  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  among the randomly initialized variants. In general, we find that increased  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  correlates with more flipped predictions in Fig. 6.

Our second key finding is illustrated in Fig. 7, which reveals that the  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  with respect to the hair color exhibits strong local fluctuations during the training for the DenseNets [28] (highest  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  difference in Fig. 6). Notably, both models classify the original sample correctly after every epoch during training. This indicates that the differences in Fig. 7 are not explained by incorrect classifications of the original image for either of the two models. Nevertheless, the networks do not continuously learn to rely on the hair color property but instead locally “forget” it, even in later epochs. This effect is particularly pronounced for the pre-trained model, whereas the randomly initialized version tends to show low  $\mathbb{E}[|\nabla_{\mathbf{x}}|]$  values. We observe similar behavior for other architectures (Appx. D.2).

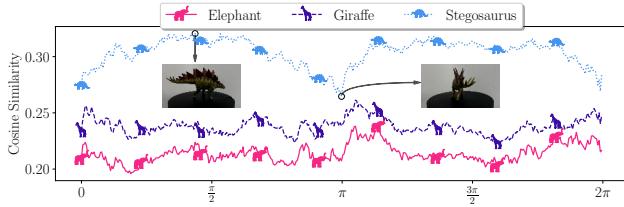


Figure 8. Average CLIP [50] cosine similarities for **real-life** interventional data. We mark high and low points for the ground truth.

#### 4.4. CLIP Zero-Shot Classification

In our final set of experiments, we investigate the widely used multimodal backbone CLIP ViT-B/32 [50] for zero-shot classification. Our approach is model-agnostic, requiring only access to model outputs, here cosine similarities in the learned latent space. Specifically, we measure the impact of interventional data by comparing ten different text descriptions against the visual embeddings. This procedure follows recent trends to increase zero-shot performance for CLIP models, e.g., [49, 56].

As the property of interest, we select object orientation, which is a known bias, for example, in ImageNet models [22]. Additionally, we demonstrate a third type of interventional data and capture real-life interventional images. Specifically, we record a rotation of three toy figures (elephant, giraffe, and stegosaurus) using a turntable. Note that during rotation, the ground truth does not change, i.e., it is the identical object. We provide the hyperparameters and additional visualizations in Appx. E.

**Results:** Fig. 8 displays the average changes in output behavior ( $\mathbb{E}[|\nabla_{\mathcal{X}}|]$  scores in Appx. E) for our real-life interventions. Specifically, we visualize zero-shot classifications of a CLIP model. Note that the behavior is remarkably consistent between text descriptors with similar standard deviations during the full interventions. Further, all measured  $\mathbb{E}[|\nabla_{\mathcal{X}}|]$  are statistically significant ( $p < 0.01$ ), i.e., the CLIP model is influenced by object orientation. While expected, e.g., [22], our local interventional approach facilitates direct interpretations of the change in behavior.

Fig. 8 reveals that the highest average similarity for the correct class occurs when the toy animal is rotated sideways. Periodical minima align with the front or back-facing orientations. In contrast, the highest similarities for the other classes appear close to the minimum of the ground truth, indicating lower confidence. In Appx. E, we include the response of CLIP to synthetic rotations of a 3D model around other axes as additional ablations. We confirm that uncommon, e.g., upside-down positions, lead to lower scores. Hence, our approach provides actionable guidance to locally select an appropriate input orientation.

## 5. Limitations

The main limitation of our work is related to a point discussed in [18]. Specifically, we rely on interventional data, which must be captured or virtually acquired. While we follow the idea of using generative models, i.e., image editing models [5, 15] together with CFG scaling [26], we understand that for these models, the causal hierarchy theorem applies [42]. However, note that [15] is trained on synthetic interventional data using [24]. Further, intervening in input space enables visual verification of whether the intervention targets the correct property. An idea related to the care set of properties in [42]. Nevertheless, we include failure cases for interventions with [15] in Appx. B.3.

Similar to [19, 32, 52], our approach is non-explorative. Hence, we need a concrete property to investigate and measure the corresponding  $\mathbb{E}[|\nabla_{\mathcal{X}}|]$ . While this has the advantage of being able to investigate unlearned properties, we believe a combination with explorative methods, e.g., [11, 30, 40, 69], is an interesting future direction.

## 6. Conclusions

By adopting a causal perspective, we study deep learning models and move beyond local associational explanations of their prediction behavior. We leverage recent breakthroughs in image-to-image editing models and Classifier-Free Guidance (CFG) scaling to gradually intervene in semantic properties. To quantify the impact of the selected properties on the predictions of a trained model, we approximate the expected property gradient magnitude  $\mathbb{E}[|\nabla_{\mathcal{X}}|]$  and verify statistical significance with a corresponding hypothesis test. Our approach offers several advantages, including the ability to locally identify causal factors and facilitate direct interpretation and quantification of corresponding output changes. To demonstrate its effectiveness, we perform an extensive empirical evaluation and study various models and tasks. First, we validate our approach on synthetically biased data and identify the causal factor before applying it to real-world skin lesion data. In both scenarios, we find that our approach outperforms local baselines in predicting locally biased behavior. Then, while investigating the training dynamics of eight classification models, we show that our  $\mathbb{E}[|\nabla_{\mathcal{X}}|]$  score locally correlates well with the number of flipped predictions. Finally, we use real-life interventions to study a pre-trained CLIP model, demonstrating that our approach can utilize diverse sources of interventional data. As black-box models continue to play an increasingly prominent role in a wide range of applications, we believe that our work can aid the development of more trustworthy systems.

**Acknowledgements:** We thank all our colleagues from the CVG Jena. In particular, Tim Büchner, Laines Schmalwasser, Gideon Stein, and Jan Blunk.

## References

- [1] International skin imaging collaboration, ISIC Archive. <https://www.isic-archive.com/>. 6, 16
- [2] Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022. 2
- [3] Jessica Bader, Leander Gurrbach, Stephan Alaniz, and Zeynep Akata. Sub: Benchmarking cbm generalization via synthetic attribute substitutions. *arXiv preprint arXiv:2507.23784*, 2025. 2
- [4] Elias Bareinboim, Juan David Correa, Duligur Ibeling, and Thomas F. Icard. On pearl’s hierarchy and the foundations of causal inference. *Probabilistic and Causal Inference*, 2022. 3, 9
- [5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 1, 3, 4, 8, 2, 5, 9
- [6] Tim Büchner, Niklas Penzel, Orlando Guntinas-Lichius, and Joachim Denzler. Facing asymmetry—uncovering the causal link between facial symmetry and expression classifiers using synthetic interventions. *arXiv preprint arXiv:2409.15927*, 2024. 2, 3, 4, 16
- [7] Tim Büchner, Niklas Penzel, Orlando Guntinas-Lichius, and Joachim Denzler. The power of properties: Uncovering the influential factors in emotion classification. *arXiv preprint arXiv:2404.07867*, 2024. 2
- [8] Chun-Hao Kingsley Chang, Elliot Creager, Anna Goldenberg, and David Kristjanson Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2018. 2, 3
- [9] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability, 2023. 1, 2
- [10] Will Cukierski. Dogs vs. cats. <https://kaggle.com/competitions/dogs-vs-cats>, 2013. Kaggle. 1, 5, 6, 3
- [11] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models, 2023. 1, 2, 8
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6, 7, 16, 17, 18, 20, 21, 22, 23
- [13] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017. 1, 2, 4
- [14] Bengt Fornberg. Generation of finite difference formulas on arbitrarily spaced grids. *Mathematics of Computation*, 51: 699–706, 1988. 4, 3
- [15] Tsu-Jui Fu, Wenzhe Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding Instruction-based Image Editing via Multimodal Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 2, 3, 4, 5, 7, 8, 6, 9, 10, 18, 19
- [16] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. 1, 2
- [17] P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer New York, 2013. 4
- [18] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019. 1, 2, 3, 4, 8
- [19] Yash Goyal, Ziyang Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019. 1, 2, 4, 8
- [20] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. 4
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7, 16, 17, 18, 20, 21, 22
- [22] Margaret Henderson and John T Serences. Biased orientation representations can be explained by experience with nonuniform training set statistics. *Journal of Vision*, 21(8): 10–10, 2021. 8
- [23] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Generating counterfactual explanations with natural language. In *ICML Workshop on Human Interpretability in Machine Learning*, pages 95–98, 2018. 2
- [24] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 8, 9
- [25] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3981–3991, 2023. 2
- [26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1, 2, 3, 4, 8, 5
- [27] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 7, 18, 20, 21, 22

- [28] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. [7](#), [18](#), [19](#), [20](#), [21](#), [22](#)
- [29] Amir-Hossein Karimi, Krikamol Muandet, Simon Kornblith, Bernhard Schölkopf, and Been Kim. On the relationship between explanation and prediction: a causal view. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. [3](#)
- [30] Tahira Kazimi, Ritika Allada, and Pinar Yanardag. Explaining in diffusion: Explaining a classifier with diffusion semantics. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 14799–14809, 2025. [1](#), [2](#), [3](#), [8](#)
- [31] Saeed Khorram and Li Fuxin. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10203–10212, 2022. [1](#), [2](#), [4](#)
- [32] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. [1](#), [2](#), [8](#)
- [33] Dimitri Korsch, Maha Shadaydeh, and Joachim Denzler. Simplified concrete dropout - improving the generation of attribution masks for fine-grained classification. In *DAGM German Conference on Pattern Recognition (DAGM-GCPR)*, 2023. [2](#)
- [34] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021. [2](#)
- [35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. [5](#)
- [36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. [7](#), [18](#), [20](#), [23](#)
- [37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. [7](#), [18](#), [19](#), [20](#), [21](#), [22](#)
- [38] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. [6](#), [7](#), [16](#), [17](#), [18](#), [20](#), [21](#), [22](#)
- [39] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. [1](#), [2](#), [5](#), [6](#), [9](#), [10](#), [11](#), [12](#)
- [40] Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Beilinkov, David Bau, and Aaron Mueller. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models, 2024. [1](#), [2](#), [8](#)
- [41] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. [1](#), [3](#)
- [42] Yushu Pan and Elias Bareinboim. Counterfactual image editing. In *Forty-first International Conference on Machine Learning*, 2024. [3](#), [4](#), [7](#), [8](#), [9](#), [16](#)
- [43] Judea Pearl. *Causality*. Cambridge University Press, 2009. [3](#), [1](#)
- [44] Karl Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. [4](#), [1](#), [3](#), [18](#), [20](#)
- [45] Niklas Penzel, Christian Reimers, Paul Bodesheim, and Joachim Denzler. Investigating neural network training on a feature level using conditional independence. In *European Conference on Computer Vision*, pages 383–399. Springer, 2022. [7](#)
- [46] Niklas Penzel, Jana Kierdorf, Ribana Roscher, and Joachim Denzler. Analyzing the behavior of cauliflower harvest-readiness models by investigating feature relevances. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 572–581. IEEE, 2023. [2](#)
- [47] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. [2](#), [3](#), [5](#)
- [48] Oana-Iuliana Popescu, Maha Shadaydeh, and Joachim Denzler. Counterfactual generation with knockoffs. *arXiv preprint arXiv:2102.00951*, 2021. [1](#), [2](#), [4](#)
- [49] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. [8](#), [27](#)
- [50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [2](#), [4](#), [8](#), [11](#), [13](#), [27](#), [29](#), [30](#), [31](#), [32](#), [33](#)
- [51] Guruprasad V Ramesh, Harrison Rosenberg, Ashish Hooda, and Kassem Fawaz. Synthetic counterfactual faces. *arXiv preprint arXiv:2407.13922*, 2024. [2](#)
- [52] Christian Reimers, Jakob Runge, and Joachim Denzler. Determining the relevance of features for deep neural networks. In *European Conference on Computer Vision*, pages 330–346. Springer, 2020. [1](#), [2](#), [3](#), [5](#), [8](#), [11](#), [13](#), [15](#)
- [53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. [1](#), [2](#), [5](#), [6](#), [9](#), [10](#), [11](#), [12](#)

- [54] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR, 2020. [6](#), [16](#)
- [55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. [3](#)
- [56] Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15746–15757, 2023. [8](#), [27](#)
- [57] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. [6](#), [7](#), [11](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#)
- [58] Adam Scherlis, Kshitij Sachan, Adam S. Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and Capacity in Neural Networks, 2023. [1](#), [2](#)
- [59] Laines Schmalwasser, Jakob Gawlikowski, Joachim Denzler, and Julia Niebling. Exploiting text-image latent spaces for the description of visual concepts. In *International Conference on Pattern Recognition (ICPR)*, 2024. (accepted at ICPR). [5](#), [2](#), [11](#), [13](#), [15](#)
- [60] Laines Schmalwasser, Niklas Penzel, Joachim Denzler, and Julia Niebling. Fastcav: Efficient computation of concept activation vectors for explaining deep neural networks. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. [2](#)
- [61] David Schuhmacher, Stephanie Schörner, Claus Küpper, Frederik Großerueschkamp, Carlo Sternemann, Celine Lugnier, Anna-Lena Kraeft, Hendrik Jütte, Andrea Tannapfel, Anke Reinacher-Schick, Klaus Gerwert, and Axel Mosig. A framework for falsifiable explanations of machine learning models with an application in computational pathology. *Medical Image Analysis*, 82:102594, 2022. [3](#)
- [62] Alon Scope, Michael A Marchetti, Ashfaq A Marghoob, Stephen W Dusza, Alan C Geller, Jaya M Satagopan, Martin A Weinstock, Marianne Berwick, and Allan C Halpern. The study of nevi in children: Principles learned and implications for melanoma diagnosis. *J. Am. Acad. Dermatol.*, 75(4):813–823, 2016. [6](#), [16](#), [17](#)
- [63] 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*, 128:336–359, 2020. [1](#), [2](#), [5](#), [6](#), [9](#), [10](#), [11](#), [12](#)
- [64] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMIR, 2017. [1](#), [2](#), [5](#), [6](#), [9](#), [11](#), [12](#)
- [65] Ilija Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021. [2](#), [4](#)
- [66] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. [1](#), [2](#), [5](#), [6](#), [9](#), [11](#), [12](#)
- [67] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. [6](#), [7](#), [16](#), [17](#), [18](#), [20](#), [21](#), [22](#)
- [68] Asher Trockman and J Zico Kolter. Patches are all you need? *Transactions on Machine Learning Research*, 2022. [5](#), [6](#), [7](#), [18](#), [19](#), [20](#), [21](#), [22](#)
- [69] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565, 2020. [1](#), [2](#), [5](#), [8](#), [11](#), [13](#), [14](#), [15](#)
- [70] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. [2](#), [3](#), [5](#), [6](#), [9](#), [10](#), [11](#), [12](#)
- [71] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *International Conference on Learning Representations*, 2017. [2](#), [3](#)