# Two-Stage Holistic and Contrastive Explanation of Image Classification


## Abstract

Explanations for the outputs of deep neural network classifiers are essential in promoting trust and comprehension among users. Conventional methods often offer explanations only for one single class in the output and neglect other classes with high probabilities, resulting in a limited view of the model's behaviors. In this paper, we propose a holistic explanation method for image classification. It not only facilitates an overall understanding of model behaviour, but also provides a framework where one can examine the evidence for discriminating competing classes, and thereby yield contrastive explanations. We demonstrate the advantages of the new method over baselines in terms of both faithfulness to the model and interpretability to users. The source code will be made available to the public upon publication of the paper.


## 1 INTRODUCTION

The output of a deep image classification model on a given input is a probability distribution over classes. This paper is concerned with the identification of pixels that are in some sense important to the classification. Previous methods such as vanilla gradient [Simonyan et al., 2014], Guided Backpropagation [Springenberg et al., 2015], LIME [Ribeiro et al., 2016], Grad-CAM [Selvaraju et al., 2017], and RISE [Petsiuk et al., 2018] focus on determining the pixels that are important for individual classes. We call them *individual-output explanation (IOX)* methods. In this paper, we advocate *whole-output explanation (WOX)* methods that explain the classification output holistically.

As a motivating example, consider the result of ResNet50 on the image in Fig. 1 (a), for which the ground-truth label is `cello`. The classes with the highest probabilities in the output are: `cello` (0.839), `acoustic-guitar` (0.081), `banjo` (0.036), `violin` (0.021), `electric-guitar` (0.008). Most previous XAI methods would explain only the top class `cello`. If Grad-CAM is used, for instance, one gets the saliency map shown in Fig. 1(d.1).

A simple way to perform WOX is to create heatmaps for each top class separately, which we call *simple WOX (SWOX)*. A well-known issue with SWOX is that it produces similar explanations for different classes [Rudin, 2019]. In Fig. 1, for instance, Grad-CAM gives essentially the same explanations for `cello` (e.1) and `violin` (e.2), as well as for `acoustic-guitar` (e.3), `banjo` (e.4) and `electric-guitar` (e.5). See Fig. 6 for more examples.

In this paper, we propose a novel method, called *contrastive WOX (CWOX)*. On the running example, it divides the top classes into two clusters: `{cello, violin}` and `{acoustic-guitar, banjo, electric-guitar}`. It performs the explanation in two steps: (1) Generate heatmaps to contrast the two clusters (Fig. 1: b, c); and (2) generate heatmaps to contrast class labels within each cluster (Fig. 1: b.1-b.2; c.1-c.3).

The heatmaps in Fig. 1 (b, c) are revealing. The evidence for Cluster 1 against cluster Cluster 2 is shown in (b), and the evidence for Cluster 2 against Cluster 1 is shown in (c). Together, the two heatmaps indicate that ResNet50 understands there are two objects. ResNet50 has some confusion as to whether the object on the right is a `cello` or a `violin`. The evidence for `cello` relative `violin` is shown in (b.1) (bottom body), and the evidence for `violin` relative to `cello` is shown in (b.2) (middle section of strings). ResNet50 also has some confusion as to whether the object on the left is an `acoustic-guitar`, a `banjo`, or an `electric-guitar`. The contrastive evidence for the three classes against each other are shown in (c.1) (bottom body), (c.2) (headstock), and (c.3) (strings) respectively. Evidently, the CWOX heatmaps are more discriminative and meaningful than the SWOX heatmaps.

Why are the top class labels grouped into the two clusters shown? The reason is that the labels from the same cluster

*Submitted to the 39th Conference on Uncertainty in Artificial Intelligence* (UAI 2023). **To be used for reviewing only**.

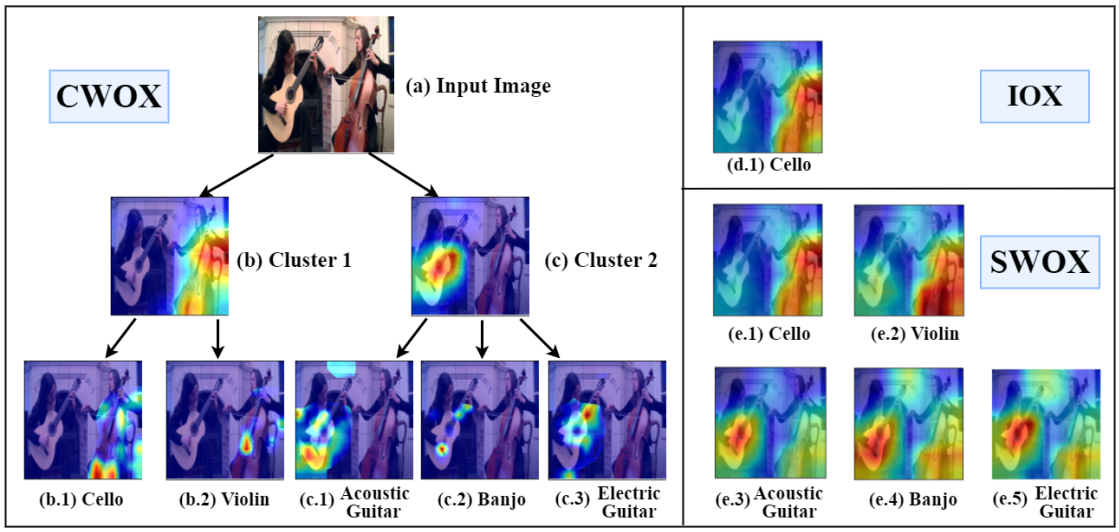



Figure 1: Three different explanations for the output of ResNet50 on the input image (a): (1) IOX explains only the top class `cello` (d.1); (2) SWOX explains all the five top classes separately (e.1 - e.5); (3) CWOX explains all the top classes in an organized and contrastive manner. It divides the top classes into two clusters, reveals the constrastive evidence that supports each cluster over the other (b, c), and does the same for the classes in each cluster (b.1-b.2, c.1-c.3).

(e.g., `cello` vs `violin`) are confusing to the classifier. The two clusters are hence called *confusion clusters*. The classes in the same cluster refer to the same object in the input image and are competing labels for that object. Thus, the first step of CWOX is essentially about *object localization* and the second step about *class discrimination*. We will discuss how to group top class labels into confusion clusters in Section 3, and how to create the heatmaps in Section 4.

In sum, IOX methods explain only one class in the output. As such, they are unable to provide users with an overall understanding of model behavior and might mislead users to unjustified confidence in the explanation method and in the model [Rudin, 2019]. This is why we need WOX methods. Between the two WOX methods, CWOX is contrastive while SWOX is not. It gives more informative and meaningful explanations than the latter. Note that contrastive explanations are easier to comprehend to humans than non-contrastive ones because humans psychologically prefer contrastive explanations [Miller, 2019].

## 2 RELATED WORK

The past few years have witnessed a surge of research activity on the explainability of deep neural networks, which is driven by the need for trust, fairness and accountability in high-stake applications [Samek et al., 2019, Li et al., 2020]. The area of eXplainable Artificial Intelligence (XAI) is consequently born. The majority of research on XAI focuses on post-hoc methods that provide explanations for the not inherently interpretable models [Samek et al., 2019, Li et al., 2020]. These methods can be further divided into *global methods*, which explain the overall behavior of a model [Kim et al., 2018, Ghorbani et al., 2019], and *local methods*, which explain the behavior of a model on a specific input [Simonyan et al., 2014, Springenberg et al., 2015, Zeiler and Fergus, 2014, Bach et al., 2015, Ribeiro et al., 2016, Selvaraju et al., 2017, Shrikumar et al., 2017, Zhang et al., 2018]. This paper focuses on post-hoc local explanations of image classification models.

The output of an image classifier is a probability distribution over multiple classes. However, most previous XAI methods aim to explain one individual output class. This only provides an understanding of a specific prediction and not the overall workings of the model [Kim and Doshi-Velez, 2021]. CWOX addresses this issue by explaining the entire output as a whole, yielding more comprehensive explanations. CWOX is a general framework and it requires an IOX method as the base explainer. In theory, any IOX method can use together with CWOX. In this sense, our work is orthogonal to most previous work.

There are previous works on contrastive explanations that involve two or more classes. Miller [2019] surveyed over 250 papers in philosophy, psychology, and cognitive science and found that humans prefer contrastive explanations that explain *why class $A$ but not class $B$* to non-contrastive ones that only explain *why $A$*. In XAI, this is achieved through either *counterfactual explanation* or *discriminative explanation*. Counterfactual explanations identify necessary modifications to change the prediction from $A$ to $B$ [Wachter et al., 2017], while discriminative explanations explicate the evidence in the input that supports $A$ over $B$ [Wang and Vasconcelos, 2020, Prabhushankar et al., 2020, Jacovi et al., 2021]. As illustrated in Fig. 1, CWOX is a systematic and organized way to conduct discriminative explanations of the top classes in classification output. Note that the naïve method that contrasts every ordered pair of top classes

| | |
|---|---|
| Cat | 1.0 |
| Rabbit | 0.0 |
| Hare | 0.0 |

| | |
|---|---|
| Cat | 0.0 |
| Rabbit | 0.6 |
| Hare | 0.4 |

| | |
|---|---|
| Cat | 0.0 |
| Rabbit | 0.3 |
| Hare | 0.7 |

Figure 2: Rabbit and hare are confusing to humans. A person would find it difficult to decide whether to label the second and third images as rabbit or hare. Similarly, a neural network classifier would give the two class labels high probabilities in either case.

would result in too many heatmaps, which might overwhelm rather than help a user.

In contrastive explanations, there is a need to identify a *contrast class (foil)* $B$ for the *target class (fact)* $A$. Previous works let the foil be: (1) all other classes (i.e., non-$A$) [Zhang et al., 2018, Jacovi et al., 2021]; (2) any other class [Dhurandhar et al., 2018, Goyal et al., 2019, Wang and Vasconcelos, 2020]; (3) the class with second highest probability [Wang and Wang, 2022]; (4) another class picked by users [Liu et al., 2019, Akula et al., 2020]; or (5) the prediction of another smaller model [Wang and Vasconcelos, 2020]. In this paper, we propose a principled method for determining how to contrast the top classes against each other. In other words, the main contribution of this paper is a way to systematically determine "foils" for contrastive explanations.

## 3 GROUPING CLASSES INTO CONFUSION CLUSTERS

As mentioned in Section 1, the notion of confusing classes is of pivotal importance to CWOX. How can we tell if two classes are confusing to a classifier? Rabbit and hare are confusing to humans. When presented with an image of either class, a person would find it difficult to decide whether to label it as a rabbit or a hare (Fig. 2). Similarly, two classes are confusing to a classifier if, when processing images containing objects of either class, it has trouble determining which of them to use as the output label. Consequently, it gives high probabilities to both classes. *It is therefore possible to determine if two classes are confusing to a classifier by checking if they often co-occur as top classes in classification outputs*.

To partition classes into confusion clusters, we first run the target classifier on a set of examples, typically the training examples. For each example, we get a list of top class labels, which we regard as a short document. For the example in Fig. 1, the document consists of five words: {`cello`, `acoustic-guitar`, `banjo`, `violin`, `electric-guitar`}. Suppose there are $N$ training examples. Then we have $N$ short documents. The task now becomes a word clustering problem. We want to partition the words (class labels) into clusters such that words from the same cluster co-occur more often in the $N$ documents than words from different clusters.

There are many methods that can be used for word clustering. We choose to use hierarchical latent tree analysis (HLTA) [Chen et al., 2017, Zhang and Poon, 2017] because it is developed specifically to model word co-occurrence in documents. We leave it to future work to evaluate other methods for this step.

HLTA is based on hierarchical latent tree models (HLTM), which are Bayesian networks with multiple levels of latent variables. An example is shown in Fig. 3. The idea is to model correlations among observed variables (the leaf nodes) using a tree of latent variables. Given a dataset on the observed variables, HLTA aims to find the model that maximizes the Bayes Information Criterion [Schwarz, 1978].

We performed HLTA on a collection of short documents obtained using ResNet50 on the training examples of ImageNet. The structure of the resulting model is partially shown in Figure 3. [1] The variables at the bottom level, level 0, are binary variables that represent the presence/absence of words in a document. The latent variables at level 1 are introduced during the analysis to model word co-occurrence patterns, e.g., $Z181$ for the co-occurrence of `hare` and `wood-rabbit`, and $Z1250$ for the co-occurrence of `mouse` and `computer-keyboard`. Latent variables at level 2 are introduced during the analysis to model the co-occurrence of the patterns at level 1, e.g., $Z29$ for the co-occurrence of the patterns $Z135$ and $Z136$.

Each node in the tree defines a cluster of class labels, which consists of the labels in the subtree rooted at the node. Some of the clusters given by level-1 nodes, for instance {`hare`, `wood-rabbit`} and {`electric-guitar`, `accoustic-guitar`, `banjo`}, consist of visually similar classes that are difficult for the classifier to discriminate. They are often competing labels for the same object/region in the input image, and hence all appear as top classes in classification output.

Class labels in some other clusters are not visually similar. One example is {`mouse`, `computer-keyboard`}. The two classes are grouped together nonetheless because mouses and keyboards tend to co-occur in images, and hence co-occur as top classes in classification output. Due to the co-occurrence, classifiers often have trouble in deciding which of them to use to label an image. If we think of a *composite object* `mouse-computer-keyboard`, then this co-occurrence cluster is no different from the visually similar clusters above: Different labels in the cluster are competing labels for the same (composite) object. See Fig. 7 for example of explanations on the composite object.

[1] The entire model structure will be posted online later.

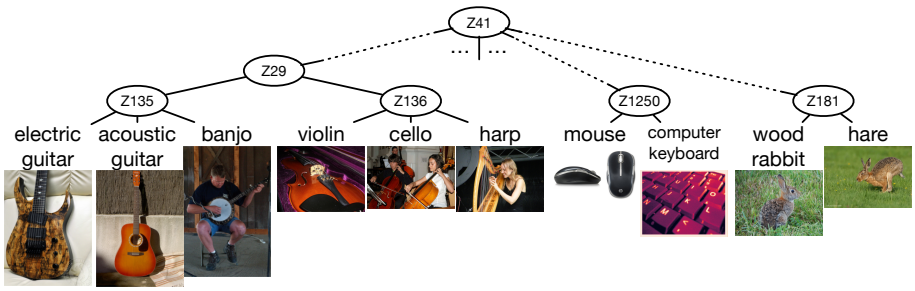



Figure 3: A part of the latent tree model built from the outputs of ResNet50 on ImageNet training examples. Solid lines are direct connections, and dashed lines are indirect connections with intermediate nodes removed. Images of the classes are displayed for visual reference. They are not part of the model. The tree reveals co-occurrence patterns of class labels in classification outputs.

There is a hierarchy behind the ImageNet classes that was derived from WordNet [Miller, 1995]. While there are some similarities, our latent tree differs from the WordNet hierarchy significantly. For example, `screwdriver` and `syringe` are far apart in WordNet, but close to each other in our latent tree due to their visual similarity. See Fig. 6 for the CWOX explanations on these two classes.

# 4 CREATING CONTRASTIVE EXPLANATIONS

Suppose we want to explain the behaviors of a classification model $m$. In our approach, the first step is to build a latent tree $T$ for all the class labels as explained in the previous section. This is done in an *offline phase*.

During the *online phase*, we create explanations for the outputs of $m$ on individual inputs. For each input image $\mathbf{x}$, we feed it to $m$ to get the $K$ top classes in the output. The value of $K$ can either be a predetermined number (e.g., 5), or the number of top classes whose total probability exceeds a threshold (e.g., 0.95).

To divide the top classes into confusion clusters, we first restrict the latent tree $T$ from the offline phase onto those classes to obtain a subtree, and then cut the subtree at level 1 to get clusters of labels. In our running example, there are five top classes: `cello`, `violin`, `acoustic-guitar`, `banjo` and `electric-guitar`. By restricting the latent tree in Fig. 3 onto those classes and cutting the resulting subtree at level 1, we get the following two clusters: `{cello, violin}` and `{acoustic-guitar, banjo, electric-guitar}`.

In general, suppose the $K$ top classes are divided into $I$ confusion clusters $\mathbf{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_I\}$, and each cluster $\mathbf{C}_i$ consists of $J_i$ class labels $\mathbf{C}_i = \{c_{i1}, \ldots, c_{iJ_i}\}$. To explain the top classes, CWOX generates a collection of *contrastive heatmaps* in two stages:

1. For each confusion cluster $\mathbf{C}_i$, create a heatmap to highlight the pixels that support $\mathbf{C}_i$ over other clusters;
2. In each $\mathbf{C}_i$, create a heatmap for each class $c_{ij}$ to highlight the pixels that support $c_{ij}$ over other classes.

## 4.1 BASE EXPLAINERS

Contrastive heatmaps are created from saliency maps for individual classes. A *saliency map* for a classes $c$ aims to highlights the pixels that are, according to the model $m$, important for the class. The more important a pixel is to the class, the higher its saliency value. It is usually computed from either the probability $P_m(c|\mathbf{x})$ or the logit $z_c(\mathbf{x})$ of the class. Saliency maps can be generated by a variety of IOX methods, including backpropagation-based techniques such as Guided Backpropagation [Springenberg et al., 2015], DeepLIFT [Shrikumar et al., 2017], Grad-CAM [Selvaraju et al., 2017]; forward propagation-based techniques like RISE [Petsiuk et al., 2018], and local approximation methods like LIME [Ribeiro et al., 2016]. They will be referred to as *base explainers* in the context of CWOX.

The concept of saliency map can easily be generalized to clusters of classes. A cluster $\mathbf{C}$ of classes can be viewed as a *compound class* with probability and logit given as follows:

$$P_m(\mathbf{C}|\mathbf{x}) = \sum_{c \in \mathbf{C}} P_m(c|\mathbf{x}), \quad z_{\mathbf{C}}(\mathbf{x}) = \log \sum_{c \in \mathbf{C}} e^{z_c(\mathbf{x})}.$$

Saliency maps can be generated for the cluster $\mathbf{C}$ in the same way as for individual classes.

## 4.2 CONTRASTIVE HEATMAPS

Let $H_{\mathbf{C}_i}$ and $H_{\mathbf{C}\setminus\mathbf{C}_i}$ be saliency maps for a confusion cluster $\mathbf{C}_i$ and all other confusion clusters respectively. $H_{\mathbf{C}_i}$ and $H_{\mathbf{C}\setminus\mathbf{C}_i}$ presumably highlight the pixels that are, according to the model $m$, important for $\mathbf{C}_i$ and $\mathbf{C}\setminus\mathbf{C}_i$ respectively. For a given pixel $x$, the difference $H_{\mathbf{C}_i}(x) - H_{\mathbf{C}\setminus\mathbf{C}_i}(x)$ measures the importance of $x$ to $\mathbf{C}_i$ relative to $\mathbf{C}\setminus\mathbf{C}_i$. Consequently, we use the following heatmap to contrast $\mathbf{C}_i$ against other confusion clusters:

$$\hat{H}_{\mathbf{C}_i} = \begin{cases} ReLU[H_{\mathbf{C}_i} - H_{\mathbf{C}\setminus\mathbf{C}_i}] & \text{if } I > 1; \\ H_{\mathbf{C}_i} & \text{if } I = 1, \end{cases} \quad (1)$$

Note that ReLU is used so as to focus on the evidence for class $\mathbf{C}_i$ rather than that against it.

Next, consider all the classes in a confusion cluster $\mathbf{C}_i$. Let $H_{c_{ij}}$ and $H_{\mathbf{C}_i \setminus c_{ij}}$ be saliency maps for a class $c_{ij} \in \mathbf{C}_i$ and all other classes in the cluster respectively. We use the following heatmap to contrast $c_{ij}$ against the other classes:

$$\hat{H}_{c_{ij}} = \begin{cases} \text{supp}(\hat{H}_{\mathbf{C}_i}) \times ReLU[H_{c_{ij}} - H_{\mathbf{C}_i \setminus c_{ij}}] \\ \text{if } J_i > 1; \\ \text{supp}(\hat{H}_{\mathbf{C}_i}, \epsilon) \times H_{c_{ij}} \\ \text{if } J_i = 1. \end{cases} \quad (2)$$

Note that the contrastive heatmap for $c_{ij}$ is restricted to $\text{supp}(\hat{H}_{\mathbf{C}_i}, \epsilon)$. This means that, when identifying contrastive evidence for classes in the $\mathbf{C}_i$, we focus only on the evidence supportive of the cluster.

An overall description of CWOX is given in Algorithm 1. As alluded to earlier, a confusion cluster $\mathbf{C}_i$ often consists of classes that are competing labels for the same region in the input image. The first step of CWOX aims to highlight that region, and hence is about *object localization*. The second step of CWOX aims to pinpoint at the evidence for each of the competing classes. It is about *class discrimination*. Class discrimination requires more fine-grained information. Some base explainers can facilitate this desiderata. In Grad-CAM, one needs to specify a pivot layer where multiple feature maps are aggregated into one heatmap using gradients from the output layer. The further away the pivot layer is from the output layer, the more fine-grained the resulting heatmap. In RISE, one needs to specify a mask size and a pixel mask probability. The resulting heatmap is more fine-grained when a smaller mask size and a lower pixel mask probability are used.

The idea of subtracting two saliency maps to create a contrastive heatmap was first proposed by [Shrikumar et al., 2017, Zhang et al., 2018]. Alternatively, one can multiply one saliency map with the "inverse" of the other [Wang and Vasconcelos, 2020]. This paper is about what contrastive heatmaps to compute rather than how to compute them.

**Algorithm 1** CWOX

I. OFFLINE PHASE

**Input**: A classification model $m$; a dataset $S$.

**Do**:

1: Feed each example $\mathbf{x}$ in $S$ to $m$ to get a list (document) of top class labels.

2: Run HLTA on the documents to get a latent tree $T$.

II. ONLINE PHASE

**Input**: A test example $\mathbf{x}$; a base explainer.

**Do**:

1: Feed $\mathbf{x}$ to $m$ to get a list of top class labels.

2: Restrict $T$ to those labels to get a subtree

3: Partition the labels into confusion clusters by cutting the subtree at level 1.

4: Create a heatmap to contrast each confusion cluster against other clusters using Equation (1).

5: In each cluster, create a heatmap to contrast each class in the cluster against other classes using Equation (2).

## 5 EMPIRICAL EVALUATIONS

In this section we evaluate CWOX against several other WOX methods that explain all top classes. The evaluations are in terms of the faithfulness and interpretability of the explanations. Here, *faithfulness* refers to an explanation's ability to accurately reflect the function learned by the model [Selvaraju et al., 2017, Petsiuk et al., 2018], while *interpretability* refers to its ability to provide a clear understanding of the relationship between input and output for human users [Ribeiro et al., 2016, Doshi-Velez and Kim, 2017].

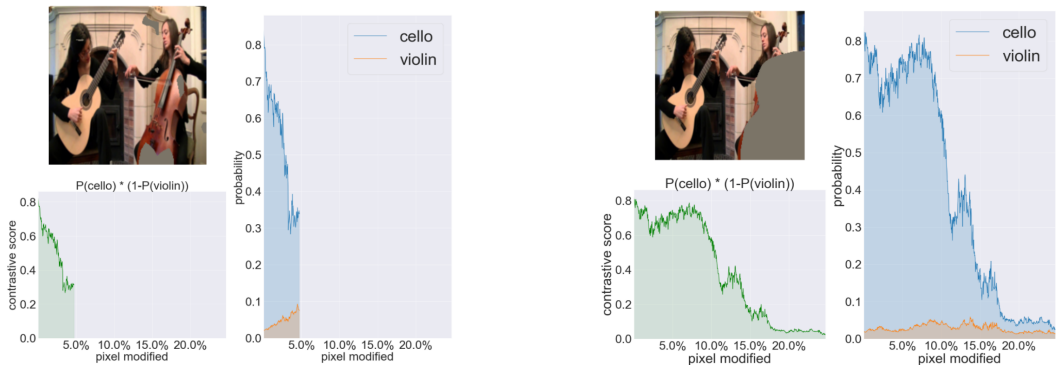
We have presented two methods, CWOX and SWOX, for explaining all top classes. A third method, CWOX-1stage (or CWOX-1s for short) is a simplified version of CWOX that contrasts each top class with all other top classes. It does not involve confusion clusters and consists of only one step. There are two possible variants. CWOX-1sA obtains a heatmap for each class by subtracting saliency maps, i.e., $H_c - H_{\mathbf{C} \setminus c}$. On the other hand, CWOX-1sB multiplies $H_c$ with the "inverse" of $H_{\mathbf{C} \setminus c}$. The second variant was proposed earlier in [Wang and Vasconcelos, 2020], where it is called SCOUT. As will be seen, the A-version of CWOX-1s significantly outperforms the B-version. Hence, we do not consider the B-version of CWOX.

To evaluate CWOX and the three baselines, we use them to explain the outputs of GoogleNet [Szegedy et al., 2015] and ResNet50 [He et al., 2016] on a subset of 10,000 images which are randomly selected from the ImageNet validation set [Deng et al., 2009]. Two base explainers are used, namely Grad-CAM [Selvaraju et al., 2017] and RISE [Petsiuk et al., 2018]. In Appendix A, we report the results with two other base explainers, as well as the setup details of the experiments.

### 5.1 FAITHFULNESS TO MODEL

**Rationale for Evaluation Metrics**: An IOX method aims to reveal the evidence a model relies on to predict a particular class. IOX explanations (saliency maps) are often evaluated in terms of their *faithfulness* to a model. Ideally, a saliency map should give high values to the pixels that the model considers important for the class. The deletion of such high-value pixels should therefore lead to a decrease in the probability of that class. This leads to a common metric for evaluating the faithfulness of an IOX explanation to a model, namely *deletion AUC* metric [Samek et al., 2016, Petsiuk et al., 2018].

Different from IOX, CWOX aims to reveal the evidence that a model uses to discriminate between classes. Consequently,




(a) CWOX: $n_\delta = 2,377$, CAUC = 0.023, CDROP = 0.505; (b) SWOX: $n_\delta = 12,402$, CAUC = 0.033, CDROP = 0.309.

Figure 4: Changes in the probabilities $P(\texttt{cello})$ and $P(\texttt{violin})$ and the contrastive score $P(\texttt{cello}) \times (1 - P(\texttt{violin}))$ as $\delta$-salient pixels are deleted according to the order induced by: (a) the CWOX heatmap (b.1) in Fig. 1, and (b) the SWOX heatmap (d.1) in Fig. 1.

Table 1: Average CAUC scores on the ImageNet examples (**smaller** ↓ CAUC indicates better contrastive faithfulness).

| | ResNet50 | | GoogleNet | |
|---|---|---|---|---|
| | Grad-CAM | RISE | Grad-CAM | RISE |
| SWOX | $7.54 \times 10^{-3}$ | $5.18 \times 10^{-3}$ | $5.93 \times 10^{-3}$ | $3.36 \times 10^{-3}$ |
| CWOX-1sA | $7.19 \times 10^{-3}$ | $4.65 \times 10^{-3}$ | $5.37 \times 10^{-3}$ | $3.12 \times 10^{-3}$ |
| CWOX-1sB | $7.68 \times 10^{-3}$ | $4.96 \times 10^{-3}$ | $6.12 \times 10^{-3}$ | $3.24 \times 10^{-3}$ |
| CWOX | $\mathbf{5.78 \times 10^{-3}}$ | $\mathbf{4.08 \times 10^{-3}}$ | $\mathbf{4.47 \times 10^{-3}}$ | $\mathbf{2.78 \times 10^{-3}}$ |

Table 2: Average CDROP scores on the ImageNet examples (**larger** ↑ CDROP indicates better contrastive faithfulness).

| | ResNet50 | | GoogleNet | |
|---|---|---|---|---|
| | Grad-CAM | RISE | Grad-CAM | RISE |
| SWOX | $6.84 \times 10^{-2}$ | $8.19 \times 10^{-2}$ | $6.56 \times 10^{-2}$ | $7.68 \times 10^{-2}$ |
| CWOX-1sA | $7.01 \times 10^{-2}$ | $8.35 \times 10^{-2}$ | $6.59 \times 10^{-2}$ | $7.73 \times 10^{-2}$ |
| CWOX-1sB | $5.22 \times 10^{-2}$ | $7.01 \times 10^{-2}$ | $5.05 \times 10^{-2}$ | $6.32 \times 10^{-2}$ |
| CWOX | $\mathbf{8.21 \times 10^{-2}}$ | $\mathbf{8.97 \times 10^{-2}}$ | $\mathbf{7.64 \times 10^{-2}}$ | $\mathbf{8.32 \times 10^{-2}}$ |

CWOX explanations should be evaluated in terms of their *contrastive faithfulness* to a model $m$, i.e., how effective they are at revealing the evidence that the model relies on to discriminate between different classes.

Suppose that a model $m$ has reasons to believe that an input $\mathbf{x}$ belongs to a class $c$, but cannot rule out the possibility of it belonging to some other classes $\mathbf{C}'$. If a heatmap $H$ is contrastively faithful to $m$, then it should give high values to the pixels that $m$ considers strongly supportive of $c$ relative to $\mathbf{C}'$. The deletion of such high-value pixels should lead to a decrease in the probability of $c$ and an increase in that of $\mathbf{C}'$. To be more specific, let there be totally $n$ pixels, and $x_1, \ldots, x_n$ be an enumeration of pixels in descending order of $H(x)$. Let $\mathbf{x}_{[r,n]}$ be the result of deleting the first $r-1$ pixels from the input image $\mathbf{x}$. If $H$ is contrastively faithful to $m$, then the probability $P_m(c|\mathbf{x}_{[r,n]})$ would decrease with $r$ and $P_m(\mathbf{C}'|\mathbf{x}_{[r,n]})$ would increase with it. Thus, the *contrastive score*: $s(r) = P_m(c|\mathbf{x}_{[r,n]})(1 - P_m(\mathbf{C}'|\mathbf{x}_{[r,n]}))$ would decrease.

As an example, consider the CWOX heatmap (b.1) in Fig. 1. It presumably reveals the evidence that ResNet50 considers supportive of `cello` relative to `violin`. Fig. 4 (a) shows what happens when pixels are deleted from the input image according to order induced by the CWOX heatmap. We see that $P(\texttt{cello})$ decreases and $P(\texttt{violin})$ increases. Consequently, the contrastive score $P(\texttt{cello}) \times (1 - P(\texttt{violin}))$ decreases. Fig. 4 (b) shows what happens when pixels are deleted according to the order induced by SWOX heatmap shown in Fig. 1 (e.1). We see that, compared to (a), the contrastive score drops more slowly at the beginning. Although a bigger drop is achieved later, it is at the expense of deleting many more pixels. Those indicate that the SWOX heatmap is less effective at pinpointing at the evidence that ResNet50 relies on to discriminate `cello` from `violin`.

**Evaluation Metrics**: We propose two quantitative metrics to evaluate the contrastive faithfulness of a heatmap to the target model. The first one is the area under the contrastive score curve, or *contrastive AUC (CAUC)* for short:

$$CAUC(H, m|\mathbf{x}, c, \mathbf{C}') = \frac{1}{n} \sum_{r=1}^{n_\delta} s(r), \qquad (3)$$

where $n_\delta$ is the number of $\delta$-*salient pixels*, i.e., the pixels with $H(x) \geq \delta \max_x H(x)$. Unlike the deletion AUC, only a subset of the pixels are included in the definition of CAUC. One reason is that we wish to focus on the most salient pixels only. Another reason is that a CWOX heatmap usually have many zero-valued pixels (See Fig. 1). There is no meaningful ordering among them. In our experiment, we set $\delta = 0.5$.

Similar to the deletion AUC, smaller CAUC scores indicate better heatmaps. In Fig. 4, the curves are shown only for the first $n_\delta$ pixels. Different heatmaps have different numbers of $\delta$-salient pixels. CAUC scores computed over different number of pixels are not comparable. Consequently, when comparing two or more heatmaps, we use the minimum of the numbers of salient pixels to calculated the CAUC scores for all of the heatmaps.

One drawback of the CAUC metric is that, when comparing two heatmaps, it does not consider all the $\delta$-salient pixels in one of the heatmaps. To consider all the $\delta$-salient pixels, we propose another metric called the *weighted drop in contrastive score (CDROP)*:

$$CDROP(H, m|\mathbf{x}, c, \mathbf{C}') = \frac{s(1) - s(n_\delta + 1)}{\log_2(1 + \frac{\max\{n_\delta, \tau\}}{\tau})}, \quad (4)$$

where $\tau$ is a hyperparameter. The score is a combination of two factors. The first factor $s(1) - s(n_\delta + 1)$ is the drop in the contrastive score due to the deletion of all the salient pixels. The second factor $\log_2(1 + \frac{\max\{n_\delta, \tau\}}{\tau})$ is a logarithmic penalty factor for $n_\delta$ when it exceeds $\tau$, which is set at $0.05n$ in our experiments (i.e., 5% of the total number of pixels). It aims to capture the intuition that too many salient pixels can be distracting to a human user, and is motivated by the Weber-Fechner law. Larger CDROP scores indicate better heatmaps.

Note that, besides the deletion AUC, the insertion AUC is often used to evaluate IOX heatmaps. We do not consider the insertion option in the contrastive setting because it is difficult to contrast different classes using a small number of pixels in the context of an "empty" canvas.

**Results**: Tab. 1 shows the CAUC scores of the four methods. They are averaged over all the 10,000 test images and, for each test image, all pairs $(c_{ij}, \mathbf{C}_i - c_{ij})$ that consist of a top class $c_{ij}$ and all the other classes from the same confusion cluster (we obtain the top classes by a probability threshold of 0.95). Tab. 2 show the corresponding CDROP scores. We see that the CAUC scores of CWOX are significantly lower than those of the baselines, and its CRDOP scores are significantly higher. Those indicate that the explanations produced by CWOX are more contrastively faithful to the models than the baselines. It is also interesting to note that CWOX-1sB is inferior to CWOX-1sA in all cases, and sometimes it is even inferior to SWOX.

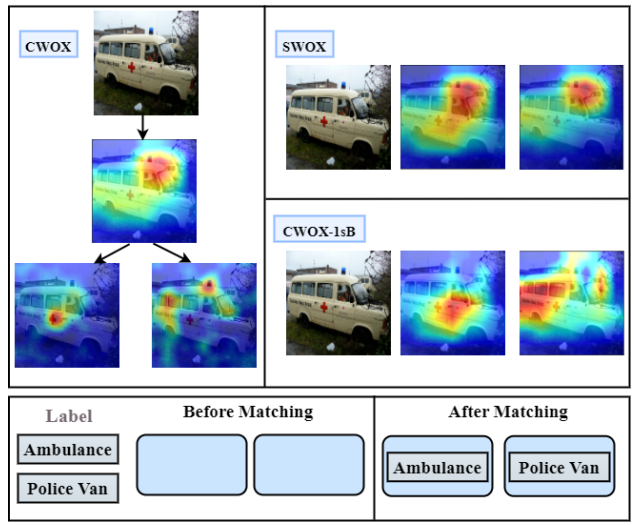



Figure 5: In the human subject study, heatmaps for pairs of confusing labels are displayed. A user is asked to match the labels with the heatmaps.

### 5.2 INTERPRETABILITY TO USERS

How well does a WOX method help human users understand the evidence that a model relies on to discriminate between classes? To answer this question we have conducted a human subject study following the *forward simulation* protocol [Ribeiro et al., 2016, Doshi-Velez and Kim, 2017, Nunes and Jannach, 2017, Lage et al., 2019, Tjoa and Guan, 2020]. As shown in Fig. 5, we display the heatmaps for pairs of confusing labels alongside the input image, and ask users to match the heatmaps with the labels. A correct matching would indicate that a user understands what pixels the model considers important for each of the two labels.

The study was conducted on the predictions by ResNet50 on a collection of images from the ImageNet validation set. For each image, a pair of confusing top classes was selected based on the latent tree $T$ from the offline phase. CWOX, CWOX-1sB and SWOX were included in the study, and RISE was used as the base explainer. To make the study manageable, we did not consider all possible combinations of models, WOX methods, and base explainers. We also limited the choices of input images and confusing class labels. See Appendix B for the details. CWOX-1sB was choosen over CWOX-1sA because it is from previous work.

60 postgraduate students were recruited for our study. At the time of the study, they were taking a course on machine learning, and were not exposed to XAI yet. They were randomly divided into three groups, and each group was responsible for only *one* of the three explanation methods. As in previous XAI user studies [Doshi-Velez and Kim, 2017, Nguyen, 2018, Hase and Bansal, 2020, Fel et al., 2021], the participants went through a training phase before handling explanations for *new unseen examples*. Besides matching

Table 3: Results of the human subject study.

| | SWOX | CWOX-1sB | CWOX |
|---|---|---|---|
| Accuracy | 0.45±0.11 | 0.57±0.20 | 0.83±0.21 |
| Confidence | 1.60±0.55 | 2.60±0.55 | 3.60±0.54 |

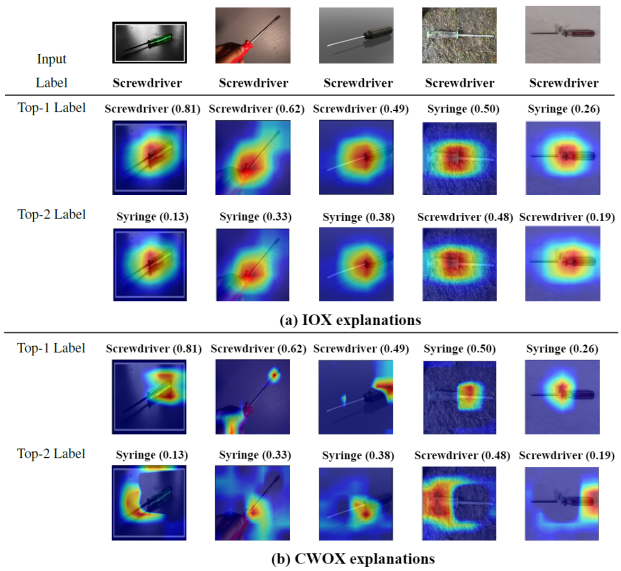



Figure 6: IOX (Grad-CAM) and CWOX explanations of the outputs of GoogleNet on five input images.

class labels with heatmaps, they were also asked to rate their confidence in their answers on a scale of 1 to 5, with 1 meaning "not sure" and 5 meaning "completely sure".

The results are shown in Tab. 3. We see that, with CWOX explanations, users were able to match the heatmaps with the class labels correctly 83% of the time, and they did it with high confidence. They did significantly worse with the explanations by SWOX and CWOX-1sB. Those results indicate that CWOX is more effective than the baselines in helping human users understand the evidence that a model relies to discriminate between classes.

To get a concrete feeling about the superiority of CWOX, imagine completing the matching tasks shown in Fig. 5. Among the two second-stage CWOX heatmaps (those at the bottom), the one on the left highlits the red cross while the one on the right does not. Hence, the former should be obviously matched with `ambulance` and the latter with `policy van`. The matching task is relatively more challenging with heatmaps by SWOW and CWOX-1sB.

### 5.3 VISUAL EXAMPLES

In this section we provide two visual examples. More examples can be found in Appendix C. The first example (Fig. 6) illustrates the differences between IOX and CWOX when used to explain the outputs of GoogleNet on five images. To a human user, the IOX explanations for the top-1 labels would suggest the same output in all the five cases. Yet, GoogleNet classifies the first three images as `screwdriver`, while the last two as `syringe`. Moreover, the explanations for the Top-2 labels are virtually the same as those for the Top-1 labels. In comparison, the CWOX explanations are more meaningful and discriminative. In fact, the CWOX heatmaps for `screwdriver` all highlight the handle, while the those for `syringe` all highlight the shaft.

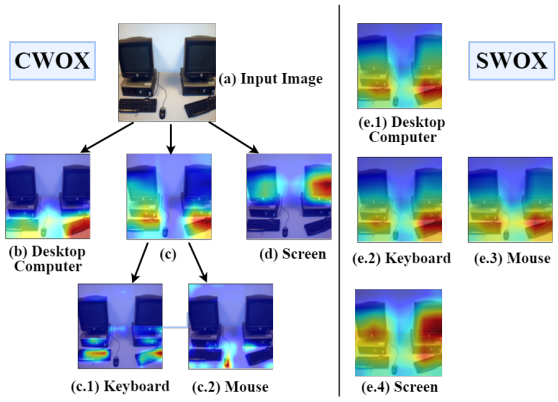



Figure 7: SWOX and CWOX explanations of the output of ResNet50 on an image that contains both keyboard and mouse.

The second example (Fig. 7) illustrates the differences between SWOX and CWOX when used to explain the output of an an image that contains both keyboard and mouse. As discussed in Section 3, `mouse` and `computer keyboard` are grouped together in the latent tree as they often co-occur in images. Consequently, CWOX first identifies evidence for the composite object `mouse-computer-keyboard` (c), and then the evidence for `keyboard` against `mouse` (c.1) and the evidence for `mouse` against `keyboard` (c.2). The CWOX explanations are clear and discriminative. However, the SWOX explanations for `mouse` and `computer keyboard` are very similar to each other, and hence are not discriminative.

## 6 CONCLUSION

We propose a novel post hoc local explanation method called CWOX for image classification. Unlike most previous methods, CWOX explains all top classes in the output rather than one individual class. The key technical contribution is a principled method for determining how to contrast the top classes against each other. The idea is to divide class labels into confusion clusters based on the frequencies they co-occur as top classes on training examples. When explaining the output on a test example, CWOX first creates heatmaps to contrast different confusion clusters, and then creates heatmaps to contrast classes in each cluster against each other. Empirical results show that, in comparison with alternative methods that explain all top classes, CWOX produces explanations that are more faithful to the model and more interpretable to human users.

## References


Arjun Akula, Shuai Wang, and Song-Chun Zhu. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2594–2601, 2020.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Peixian Chen, Nevin L Zhang, Tengfei Liu, Leonard KM Poon, Zhourong Chen, and Farhan Khawar. Latent tree models for hierarchical topic detection. *Artificial Intelligence*, 250:105–124, 2017.

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

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pages 592–603, 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Thomas Fel, Julien Colin, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *arXiv preprint arXiv:2112.04417*, 2021.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pages 9277–9286, 2019.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019.

Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*, 2020.

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.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Contrastive explanations for model interpretability. *arXiv preprint arXiv:2103.01378*, 2021.

Been Kim and Finale Doshi-Velez. Machine learning techniques for accountability. *AI Magazine*, 42(1):47–52, 2021.

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, 2018.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*, 2019.

Xiao-Hui Li, Caleb Chen Cao, Yuhan Shi, Wei Bai, Han Gao, Luyu Qiu, Cong Wang, Yuanyuan Gao, Shenjia Zhang, Xun Xue, et al. A survey of data-driven and knowledge-aware eXplainable AI. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative counterfactual introspection for explainable deep learning. *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5, 2019.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, 2018.

Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3):393–444, 2017.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.

Mohit Prabhushankar, Gukyeong Kwon, Dogancan Temel, and Ghassan AlRegib. Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3289–3293. IEEE, 2020.

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.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

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. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3145–3153, 2017.

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

J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. URL `http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a`.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32 (11):4793–4813, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

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

Yipei Wang and Xiaoqian Wang. "why not other classes?": Towards class-contrastive back-propagation explanations. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=X5eFS09r9hm`.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

Nevin L Zhang and Leonard KM Poon. Latent tree analysis. *AAAI*, pages 4891–4898, 2017.