Explanations for Occluded Images

Hana Chockler
causaLens and
King's College London
hana.chockler@kcl.ac.uk

Daniel Kroening*
Amazon.com, Inc.
daniel.kroening@gmail.com

Youcheng Sun
Queen's University Belfast
youcheng.sun@qub.ac.uk

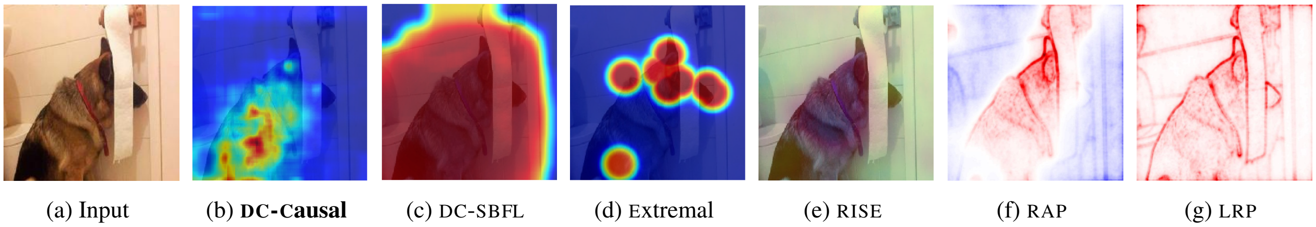

(a) Input (b) **DC-Causal** (c) DC-SBFL (d) Extremal (e) RISE (f) RAP (g) LRP

Figure 1: Explanations for the classification as 'German shepherd dog' (a). DC-Causal is the tool presented here.

Abstract

Existing algorithms for explaining the output of image classifiers perform poorly on inputs where the object of interest is partially occluded. We present a novel, black-box algorithm for computing explanations that uses a principled approach based on causal theory. We have implemented the method in the DeepCover *tool. We obtain explanations that are much more accurate than those generated by the existing explanation tools on images with occlusions and observe a level of performance comparable to the state of the art when explaining images without occlusions.*

1. Introduction

Deep neural networks (DNNs) are now a primary building block of many computer vision systems. DNNs are complex non-linear functions with algorithmically generated (and not engineered) coefficients. In contrast to traditionally engineered image processing pipelines it is difficult to retrace how the pixel data are interpreted by the layers of the DNN. This "black box" nature of DNNs creates demand for techniques that explain why a particular input yields the output that is observed.

An explanation of an output of an automated procedure is essential in many areas, including verification, planning, diagnosis and the like. A good explanation can increase a user's confidence in the result. Explanations are also useful for determining whether there is a fault in the automated procedure: if the explanation does not make sense, it may indicate that the procedure is faulty. It is less clear how to define what a *good* explanation is. There have been a number of definitions of explanations over the years in various domains of computer science [3, 11, 23], philosophy [16] and statistics [26]. The recent increase in the number of machine learning applications and the advances in deep learning led to the need for *explainable AI*, which is advocated, among others, by DARPA [12] to promote understanding, trust, and adoption of future autonomous systems based on learning algorithms (and, in particular, image classification DNNs). DARPA provides a list of questions that a good explanation should answer and an epistemic state of the user after receiving a good explanation. The description of this epistemic state boils down to *adding useful information* about the output of the algorithm and *increasing trust* of the user in the algorithm.

Explanations for the results of image classifiers are typically based on or are given in the form of a *ranking* of the pixels, which is a numerical measure of importance: the higher the score, the more important the pixel is for the DNN's classification outcome.

A user-friendly explanation can then be a subset of highest-ranked pixels that is sufficient for the original classification outcome. Given an image that features an object, good algorithms are able to generate rankings that identify that object with a high accuracy. Another typical proxy for the quality of a ranking is how many of the high-ranked pixels (constituting an explanation) need to be masked before the classification generated by the DNN changes. Good

*The work reported in this paper was done prior to joining Amazon.

explanations require very little masking.

In the absence of further analysis, the space of possible orderings (and hence rankings) is exponential in the size of the image, and the brute-force approach is therefore impractical. Moreover, as we argue in this paper, the problem is NP-complete. It is therefore expected that explanation-generating algorithms approximate the solution using heuristics tuned for image classification. This assumption is entirely appropriate in many use cases, and in particular, works very well on the benchmark sets that are used in the area: the existing work has been evaluated using the ImageNet dataset and ImageNet has been curated so are all objects clearly visible. Consequently, the explanations that are generated are usually contiguous.

We argue that there is a use-case for explanations of the results of image classifiers for images where the trigger for the result is *not* contiguous. Obvious exemplars are images with partial occlusion, say by a person walking into a scene or simply by dirt on your camera lens. To quantify the quality of the explanations for such images objectively, we introduce a new image dataset we call *Photo Bombing*, in which we obscure ImageNet photos by masking parts of the object. The difference between the modified image and the original one is the ground truth for the "photobomber", and a good explanation has no overlap with it.

We observe that the existing methods for explaining the outcome of image classifiers perform poorly on such inputs. To address this problem, we introduce a new algorithm that is grounded in causal theory. Our algorithm is iterative and highly parallelizable and delivers significantly better accuracy on an existing dataset with partial occlusion and on our own photo bombing data set. The tool, the new benchmark set, and the full set of results are available at https://www.cprover.org/deepcover/.

2. Related Work

There is a large body of work on explaining image classifiers. The existing approaches can be largely grouped into two categories: propagation and perturbation.

Propagation-based explanation methods are often regarded as more efficient. They back-propagate a model's decision to the input layer to determine the weight of each input feature for the decision. Grad-CAM [27] only needs one backward pass and propagates the class-specific gradient into the final convolutional layer of a DNN to coarsely highlight important regions of an input image. Guided Back-Propagation (GBP) [29] computes the single and average partial derivatives of the output to attribute the prediction of a DNN. Integrated Gradients (IG) [31] further uses two axioms called sensitivity and implementation invariance for the problem of how to attribute the classification by a deep network to its input features. Layer-wise relevance propagation (LRP) [1] is defined by a set of constraints on the layers and it assumes that the classifier can be decomposed into several layers of computation. In [28], the activation level of each neuron is compared with some reference point, and its contribution score for the final output is assigned according to the difference. RAP (Relative Attributing Propagation) [22] attributes a positive and negative relevance to each neuron, according to its relative influence among the neurons. SHAP (SHapley Additive exPlanations) [21] goes beyond propagation and identifies inputs that are similar to the input for which the output is to be explained, and ranks the features of the input according to their difference. A key advantage of SHAP is that it does not require counterfactuals.

In contrast to propagation-based explanation methods, perturbation-based explanation approaches explore the input space directly in search for an explanation. The exploration/search often requires a large number of inference passes, which incurs significant computational cost when compared to propagation methods. Many sampling methods have been proposed, but most of are based on random search or heuristics and lack rigor.

Given a particular input, LIME [25] samples the the neighborhood of this input and creates a linear model to approximate the system's local behavior; owing to the high computational cost of this approach, the ranking uses super-pixels instead of individual pixels. In [6], the natural distribution of the input is replaced by a user-defined distribution and the Shapley Value method is used to analyze combinations of input features and to rank their importance. In [4], the importance of input features is estimated by measuring the the flow of information between inputs and outputs. Both the Shapley Value and the information-theoretic approaches are computationally expensive. In RISE [24], the importance of a pixel is computed as the expectation over all local perturbations conditioned on the event that the pixel is observed. The concept of "extreme perturbations" has been introduced to improve the perturbation analysis by the Extremal algorithm [10]. More recently, spectrum-based fault localisation (SBFL) has been applied to explaining image classifiers. The technique has been implemented in the tool DeepCover [30], and we refer to it as DC-SBFL. It outperforms the other tools when explaining images without occlusions. Both RISE and DC-SBFL mask input pixels randomly. By contrast, Extremal uses an area constraint to optimize the perturbation. Like our new method, DC-SBFL constructs explanations greedily from a ranked list of pixels. The ranking, however, is calculated with statistical fault localisation, and is much less precise, as we demonstrate empirically.

The work presented in this paper is motivated by the fact that compositionality is a fundamental aspect of human cognition [2, 9] and by the compositional computer vision work in recent years [19, 32, 19, 35, 33, 34]. While it is well known that the performance of conventional convolutional neural networks degrades when given partially occluded objects, the impact of partial occlusion on algorithms for generating explanations has not been studied before.

The DC-Causal method presented in this paper is a perturbation-based approach and addresses the limitations of existing black-box methods in two aspects. The feature masking in DC-Causal uses causal reasoning that provides a guarantee (subject to the assumption). Furthermore, the DC-Causal algorithm is highly parallelizable, which makes it ideal for large-scale computer vision problems. As we demonstrate in Section 4.4, DC-Causal constructs its explanations in a compositional manner and is therefore a perfect fit for compositional computer vision pipelines.

3. Theoretical Foundations

In this section we describe the theoretical foundations of our approach.

3.1. Background on Actual Causality

Our definitions are based on the framework of *actual causality* introduced in [14]. Due to the lack of space, we do not present the framework here in full, but instead discuss the intuition informally. The definition of an *actual cause* is based on the concept of *causal models*, which consists of the set of variables, the range of each variable, and the structural equations describing the dependencies between the variables. Actual causes are defined with respect to a given causal model, a given context (an assignment to the variables of the model), and a propositional logic formula that holds in the model in this context.

Actual causality extends the simple counterfactual reasoning [17] by considering *contingencies*, which are changes of the current setting. Roughly speaking, a subset of variables X and their values in a given context is an actual cause of a Boolean formula φ being True if there exists a change in the values of other values that creates a counterfactual dependency between the values of X and φ (that is, if we change the values of variables in X, φ would be falsified). The formal definition is more complex and requires that the dependency is not affected by changing the values of variables not in the contingency, as well as requesting minimality.[1]

Responsibility, defined in [5], is a quantification of causality, attributing to each actual cause its *degree of responsibility*, which is based on the size of a smallest contingency required to create a counterfactual dependence. Essentially, the degree of responsibility is defined as $1/(k+1)$, where k is the size of a smallest contingency. The degree of responsibility of counterfactual causes is therefore 1 (as $k=0$), and the degree of responsibility of variables that have no causal influence on φ is 0, as k is taken to be ∞. In general, the degree of responsibility is always between 0 and 1, with higher values indicating a stronger causal dependency.

3.2. Causes in image classification

Our definition of causality for image classification follows the contingency-based approach and is derived from the definition in [14] and its matching definition of responsibility [5]; the variables represent the pixels of an input image. We cannot assume any dependency between the pixels of the image, hence we do not define any structural equations on the variables. As our goal is ranking of the pixels according to their importance for the classification, we only consider *singleton causes.*

Definition 1 (Singleton cause for image classification)**.** *For an image x classified by the DNN as $f(x)=o$, a pixel p_i of x is a* cause *of o iff there exists a subset P_j of pixels of x such that the following conditions hold:*

SC1. *$p_i \notin P_j$;*

SC2. *changing the color of any subset $P_j' \subseteq P_j$ to the masking color does not change the classification;*

SC3. *changing the color of P_j and the color of p_i to the masking color changes the classification.*

We call such P_j a witness *to the fact that p_i is a cause of x being classified as o.*

Definition 2 (Simplified responsibility)**.** *The* degree of responsibility *$r(p_i,x,o)$ of p_i for x being classified as o is defined as $1/(k+1)$, where k is the size of the smallest witness set P_j for p_i. If p_i is not a cause, k is defined as ∞, and hence $r(p_i,x,o)=0$. If changing the color of p_i alone to the masking color results in a change in the classification, we have $P_j=\emptyset$, and hence $r(p_i,x,o)=1$.*

Lemma 1. *Definition 1 is equivalent to the definition of actual cause when all variables in the model are independent of each other.*

Proof sketch. The definition matches the definition of a singleton cause for binary causal models, where a variable has the value 1 if the corresponding pixel is set to its original color, and 0 if the pixel is set to the masking color. The value of φ is 1 for the original classification (and 0 otherwise). The *context* assigns all variables the value 1, corresponding to the original image; hence φ has the value 1. The minimality requirement is satisfied immediately, given that our causes are singletons. The additional condition in [14] that requires subsets Z to be set to their original values is only relevant when there are dependencies between the variables. □

Corollary 1. *The problem of detecting causes in image classification is NP-complete.*

This result follows from Lemma 1 and [7].

[1] In [13], Halpern presents an updated definition of causality; the version in [14] is more suitable for our purposes, as we are interested in causes consisting of a single element.

Observation 1. *Given an image x and its classification o, we can calculate the degree of responsibility of each pixel p_i of x by directly applying Def. 1, that is, by checking the conditions* **SC1***,* **SC2***, and* **SC3** *for all subsets P_j of pixels of x and then choosing a smallest witness subset. While there is an underlying Boolean formula that determines the classification o given the values of the pixels of x, we do not need to discover this formula in order to calculate the degree of responsibility of each pixel of x.*

3.3. Explanations in image classification

We adapt the definition of explanations by Halpern and Pearl [15] to our setting. The definition in [14] is derived from the definition of actual causality. Our definition is based on Def. 1.

Definition 3 (Explanation for image classification)**.** *An explanation in image classification is a minimal subset of pixels of a given input image that is sufficient for the DNN to classify the image, where "sufficient" is defined as containing only this subset of pixels from the original image, with the other pixels set to the masking color.*

We note that (1) the explanation cannot be too small (or empty), as a too small subset of pixels would violate the sufficiency requirement, and (2) there can be multiple explanations for a given input image.

The precise computation of an explanation in our setting is intractable, as the problem is equivalent to the earlier definition of explanations in binary causal models, which is DP-complete [8] (the proof is similar to the proof of Lemma 1). The brute-force approach of checking the effect of changing the color of each subset of pixels of the input image to the masking color is exponential in the size of the image. Instead, we introduce a *greedy compositional approach* to computing explanations. The approach is greedy because we rank the elements in the decreasing order of responsibility for the classification and greedily add them to the explanation one by one until the original classification is restored. This approach generates explanations that are minimal in the sense that no pixels can be removed from them without altering the classification; however, they are not necessarily minimal in size, and hence are, strictly speaking, an *approximation* of Def. 3.

Unfortunately, extracting approximate explanations from the ranking is still intractable, as the ranking is based on computing the degree of responsibility of each pixels, which is NP-complete[2].

In the next section we introduce a compositional approach to computing the (approximate) degree of responsibility. The approach is based on the notion of a *super-pixel* P_i, which is a subset of pixels of a given image. Given an image x, we partition it into a small number of superpixels and compute their degree of responsibility for the output of the DNN. Then, we only refine those superpixels with a high responsibility (exceeding a predefined threshold). The scalability of the approach relies on the following observation, which is heuristically true in our experiments.

Observation 2. *The pixels with the highest responsibility for the DNN's decision are located in super-pixels with the highest responsibility.*

Intuitively, the observation holds when pixels with high responsibility do not appear in the superpixels surrounded by other pixels with very low responsibility for the input image classification outcome. While this can happen in principle, we do not encouter this case in practice owing to the continuous nature of images (even when the explanation is non-contiguous). This property is key to the success of our algorithm.

4. Compositional Explanations

In this section, we present our *greedy compositional explanation (CE)* algorithm. The general idea is to calculate the responsibility of a superpixel and recursively distribute this responsibility to all pixels within this superpixel. The CE approach in this work consists of three steps.

1. Given a set of superpixels, compute the responsibility of each its superpixel (Section 4.1).

2. Following the responsibility result in Step 1, further refine the superpixel and calculate the responsibility for the refined superpixels (Section 4.2).

3. As it is insufficient to explain an input by only using one particular set of superpixels, multiple sets will be selected and they will be analysed independently by Step 2. Finally, all their results will be merged and a ranking of pixels following their responsibility will be computed, from which an explanation will be constructed (Section 4.3).

Section 4.4 gives a step-by-step example to illustrate the working of the algorithm.

4.1. Computing the responsibility of a superpixel

Given a set of pixels $\mathcal{P}$, we use $\mathbb{P}_i$ to denote a *partition* of $\mathcal{P}$, that is, a set $\{P_{i,j} : \bigcup P_{i,j} = \mathcal{P} \text{ and } \forall j \neq k, P_{i,j} \cap P_{i,k} = \emptyset\}$. The number of elements in $\mathbb{P}_i$ is a parameter, denoted by s; in this work, we consider $s = 4$. We refer to $P_{i,j}$ as *superpixels*.

For a DNN $\mathcal{N}$, an input x, and a partition $\mathbb{P}_i$, we can generalize Def. 1 to the set of *superpixels* defined by $\mathbb{P}_i$. We denote by $r_i(P_{i,j}, x, \mathcal{N}(x))$ the *degree of responsibility* of a superpixel $P_{i,j}$ for $\mathcal{N}$'s classification of x, given $\mathbb{P}_i$.

[2] The decision problem is NP-complete; the corresponding function problem is $\text{FP}^{\text{NP}[\log n]}$-complete.

For a partition $\mathbb{P}_i$, we denote by X_i the set of *mutant images* obtained from x by masking subsets of $\mathbb{P}_i$, and by $\tilde{X}_i$ the subset of X_i that is classified as the original image x. Formally,

$$\tilde{X}_i = \{x_m : \mathcal{N}(x_m) = \mathcal{N}(x)\}.$$

We compute $r_i(P_{i,j}, x, \mathcal{N}(x))$, the responsibility of each superpixel $P_{i,j}$ in the classification of x, in Alg. 1. For a superpixel $P_{i,j}$, we define the set

$$\tilde{X}_i^j = \{x_m : P_{i,j} \text{ is not masked in } x_m\} \cap \tilde{X}_i.$$

For a mutant image x_m, we define $\mathit{diff}_i(x_m, x)$ as the number of superpixels in the partition $\mathbb{P}_i$ that are masked in x_m (that is, the difference between x and x_m with respect to $\mathbb{P}_i$). For an image y, we denote by $y(P_{i,j})$ an image that is obtained by masking the superpixel $P_{i,j}$ in y.

The degree of responsibility of a superpixel $P_{i,j}$ is calculated by Alg. 1 as a minimum difference between a mutant image and the original image over all mutant images x_m that do not mask $P_{i,j}$, are classified the same as the original image x, and masking $P_{i,j}$ in x_m changes the classification.

Algorithm 1 *responsibility*$(x, \mathbb{P}_i)$

```
INPUT: an image $x$, a partition $\mathbb{P}_i$
OUTPUT: a responsibility map $\mathbb{P}_i \to \mathbb{Q}$

for each $P_{i,j} \in \mathbb{P}_i$ do
    $k \leftarrow \min_{x_m}\{\mathit{diff}(x_m, x) \mid x_m \in \tilde{X}_i^j\}$
    $r_{i,j} \leftarrow \frac{1}{k+1}$
end for
return $r_{i,0}, \ldots, r_{i,|P_i|-1}$
```

4.2. Compositional refinement of the responsibility

Alg. 1 calculates the responsibility of each superpixel, subject to a given partition. Then, it proceeds with only the high-responsibility superpixels. Note that in general, it is possible that all superpixels in a given partition have the same responsibility. Consider, for example, a situation where the explanation is right in the middle of the image, and our partition divides the image into four quadrants. Each quadrant would be equally important for the classification, hence we would not gain any insight into why the image was classified in that particular way. In this case, the algorithm chooses another partition.

Our compositional algorithm (see Alg. 2) iteratively refines the high-responsibility superpixels until a precise explanation is constructed and recursively applies Alg. 1 to each refinement.

Given a partition, Alg. 2 calculates the responsibility for each superpixel (Line 1). If the termination condition is met (Lines 2–3), the responsibility map Q is updated accordingly. Otherwise, for each superpixel in $\mathbb{P}_i$ with responsibility higher than 0, we refine it and call the algorithm recursively (0 is a parameter that can be replaced with a sufficiently low threshold without affecting the quality of the explanation; $\mathbb{P}_{i,j}$ is an arbitrary partition of the superpixel $P_{i,j}$). We use $\cup$ to include these newly computed values in the returned map. The algorithm terminates when: 1) the superpixels in $\mathbb{P}_i$ are sufficiently refined (containing only very few pixels), or 2) when all superpixels in $\mathbb{P}_i$ have the same responsibility (this condition is for efficiency).

Algorithm 2 *compositional_responsibility*$(x, \mathbb{P}_i)$

```
INPUT: an image $x$ and a partition $\mathbb{P}_i$
OUTPUT: a responsibility map $\mathbb{P}_i \longrightarrow \mathbb{Q}$

1: $R \leftarrow$ responsibility$(x, \mathbb{P}_i)$
2: if $R$ meets termination condition then
3:     return $R$
4: end if
5: $R' \leftarrow \emptyset$
6: for each $P_{i,j} \in \mathbb{P}_i$ s.t. $R(P_{i,j}) \neq 0$ do
7:     $R' \leftarrow R' \cup$ compositional_resposibility$(x, \mathbb{P}_{i,j})$
8: end for
9: return $R'$
```

4.3. Compositional explanation algorithm

So far, we assume one particular partition $\mathbb{P}_i$, which Alg. 2 recursively refines and calculates the corresponding responsibilities of superpixels in each step by calling Alg. 1. We note that the choice of the initial partition can affect the values calculated by Alg. 2, as the partition determines the set of mutants in Alg. 1. We ameliorate the influence of the choice of any particular partition by iterating the algorithm over a set of partitions. In Alg. 3, we consider N partitions and compute an average of the degrees of responsibility induced by each of these partitions. In the algorithm, $\mathbb{P}^x$ stands for a specific partition chosen randomly from the set of partitions, and r_p denotes the degree of responsibility of a pixel p w.r.t. $\mathbb{P}^x$.

Alg. 3 has two parts: ranking all pixels (Lines 1–9) and constructing the explanation (Lines 10–17). The algorithm ranks the pixels of the image according to their responsibility for the model's output. Each time a partition is randomly selected (Line 3), the compositional refinement (Alg. 2) is called to refine it into a set of fine-grained superpixels and calculate their responsibilities (Line 4). A superpixel's responsibility is evenly distributed to all its pixels, and the pixel-level responsibility is updated accordingly for each sampled partition (Lines 5–7). After N iterations, all pixels are ranked according to their responsibility r_p.

Algorithm 3 *compositional_explanation*(x)

```
INPUT: an input image $x$, a parameter $N \in \mathbb{N}$
OUTPUT: an explanation $\mathcal{E}$

$r_p \leftarrow 0$ for all pixels $p$
for $c$ in 1 to $N$ do
    $\mathbb{P}^x \leftarrow$ sample a partition
    $R \leftarrow compositional\_responsibility(x, \mathbb{P}^x)$
    for each $P_{i,j} \in$ domain of $R$ do
        $\forall p \in P_{i,j} : r_p \leftarrow r_p + \frac{R(P_{i,j})}{|P_{i,j}|}$
    end for
end for
$pixel\_ranking \leftarrow$ pixels from high $r_p$ to low
$\mathcal{E} \leftarrow \emptyset$
for each pixel $p_i \in pixel\_ranking$ do
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{p_i\}$
    $x^{exp} \leftarrow$ mask pixels of $x$ that are not in $\mathcal{E}$
    if $\mathcal{N}(x^{exp}) = \mathcal{N}(x)$ then
        return $\mathcal{E}$
    end if
end for
```

The remainder of Alg. 3 follows the method for explaining the result of an image classifier in [30]. That is, we construct a subset of pixels $\mathcal{E}$ to explain $\mathcal{N}$'s output on this particular input x *greedily*. We add pixels to $\mathcal{E}$ as long as $\mathcal{N}$'s output on $\mathcal{E}$ does not match $\mathcal{N}(x)$. This process terminates when $\mathcal{N}$'s output is the same as on the whole image x. The set $\mathcal{E}$ is returned as an explanation.

While we approximate the computation of an explanation in order to ensure efficiency of our approach, the algorithm is built on solid theoretical foundations, which distinguishes it from other random or heuristic-based approaches. In practice, while our algorithm uses an iterative average of a greedy approximation, it yields highly accurate results. The tool website features a comparison with the exact computation of explanations for small images of 4x4 pixels, showing that DC-Causal's explanations are optimal in more than 96% of the cases. Furthermore, our approach is simple and general, and uses the DNN as a blackbox. Finally, an important advantage of our algorithm is its high parallelizability, as each partition can be analyzed in parallel. This opens an opportunity for further performance improvements.

4.4. Illustrative example

To illustrate how DC-Causal works consider Figure 2a, which is classified as 'bus' by the compositional net [18], even though there is an occlusion in the middle. Initially, DC-Causal picks up an arbitrary partition of the image with four superpixels, as in Figure 2b. This results in 15 combinations of masking superpixels, which are given to Alg. 1 to calculate the responsibility of each superpixel. The refinement of a superpixel happens in Alg. 2. As in Figure 2c, an initial superpixel is further partitioned into four more fine-grained superpixels. Overall, the refinement done by Alg. 2 for this particular example goes into two levels (Line 7 in Alg. 2) before the stopping condition is reached (Line 2 in Alg. 2). The heat-map in Figure 3 (a) gives the importance of each pixel hen starting from the single partition in Figure 2b. Though DC-Causal still highlights the important feature in the front of the bus, the result is coarse. As the number of iterations down by DC-Causal (Alg. 3) increases and a larger number of initial partitions is sampled, the result quickly converges (Figure 3), and the important features are identified successfully, avoiding the occlusion in the input image. The final explanation found by DC-Causal is given as Figure 4.

4.5. Complexity and comparison with existing work

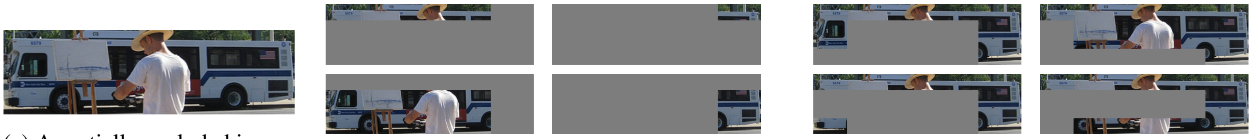
We argue that our approach is more efficient than other explanation-generating approaches to image classification. We first discuss the theoretical complexity of Alg. 3. Our claims are further supported by the empirical results presented in Sec. 5.

The SHAP (SHapley Additive exPlanations) method [21] unifies the theory behind several popular methods for explaining AI models. Assuming feature independence, the SHAP approach has runtime complexity $O(2^M + M^3)$, where M is the is the number of input features, bounded by the number of pixels of the input image, and is hence exponential in the size of the input image. Even if in practice we rarely observe exponential running time, we stress that there is no guarantee that SHAP finds the important features of the image in reasonable time. The problem becomes even harder when the important features are distributed in multiple, non-contiguous parts of an input image.

By contrast, our algorithm is based on causal theory and is inherently modular. Furthermore, the runtime complexity of our algorithm is at most *linear* in the size of image, as the following lemma proves.

Lemma 2. *The runtime of Alg. 3 is $O(2^s nN)$, where s is the size of the partition in each step (in our setting $s = 4$), n is the number of pixels in the original image x, and N is the number of initial partitions.*

Proof sketch. The computation of responsibilities of superpixels in one partition is $O(2^s)$, as the algorithm examines the effect of mutating each subset of the superpixels in the current partition. Note that s is a constant independent of the size of the image. The number of steps is determined by the termination condition on the size of a single superpixel. In our setting, the algorithm terminates when a single superpixel is $1/10$ of the original image, thus resulting in a constant-time computation. However, in general, the algo-

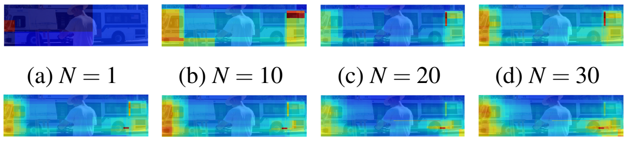


(a) A partially occluded image that is classified as 'bus' by the compositional net [18]

(b) The initial four superpixels chosen by DC-Causal (Line 3 in Alg. 3)

(c) Further refinement of the (lower left) superpixel in Figure 2b (Line 7 in Alg. 2)

Figure 2: An illustrative example using DC-Causal to explain an image from the partial occlusion image data set [19, 32]



(a) $N = 1$ (b) $N = 10$ (c) $N = 20$ (d) $N = 30$

(e) $N = 40$ (f) $N = 50$ (g) $N = 70$ (h) $N = 100$

Figure 3: Improvement of the DC-Causal's pixel ranking as the number of initial partitions N increases (Alg. 3)

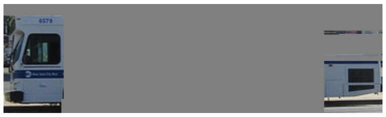
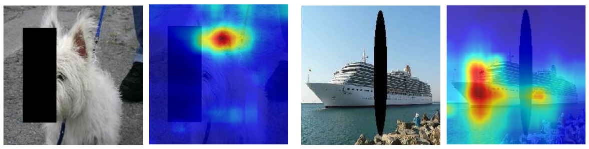


Figure 4: Explanation found by DC-Causal for Fig. 2a for the ranking in Fig. 3f



(a) 'West Highland white terrier' (b) 'Ocean liner'

Figure 5: Photo Bombing images and output from DC-Causal

rithm can continue down to the level of a single pixel, thus resulting in n pixels in the last step, hence the factor n. The algorithm performs N iterations, and every iteration uses a different initial partition. □

5. Evaluation

Benchmarks and Setup We have implemented the proposed compositional explanation approach in the publicly available tool DeepCover[3]. In the evaluation, we compare DC-Causal with a wide range of explanation tools, including the most recent ones and popular ones like DC-SBFL [30], Extremal [10], RISE [24], RAP [22], LRP [1], IG [31] and GBP [29]. We test both the Compositional Net that has been designed for partially occluded images [18] and standard convolutional models for ImageNet. We use three data sets: images known to feature partial occlusion [19, 32], a "Photobombing" data set with added occlusions for with we have ground truth, and the "Roaming Panda" data set, which contains modified ImageNet images without occlusion [30].

There is no single best way to evaluate the quality of an explanation. In this work, we quantify the quality of the explanations with two complementary methods: 1) the explanation size, and 2) the intersection with the planted occlusion part of an image. Intuitively, a good explanation should be a part of the original input and it should not intersect much with the occlusion.

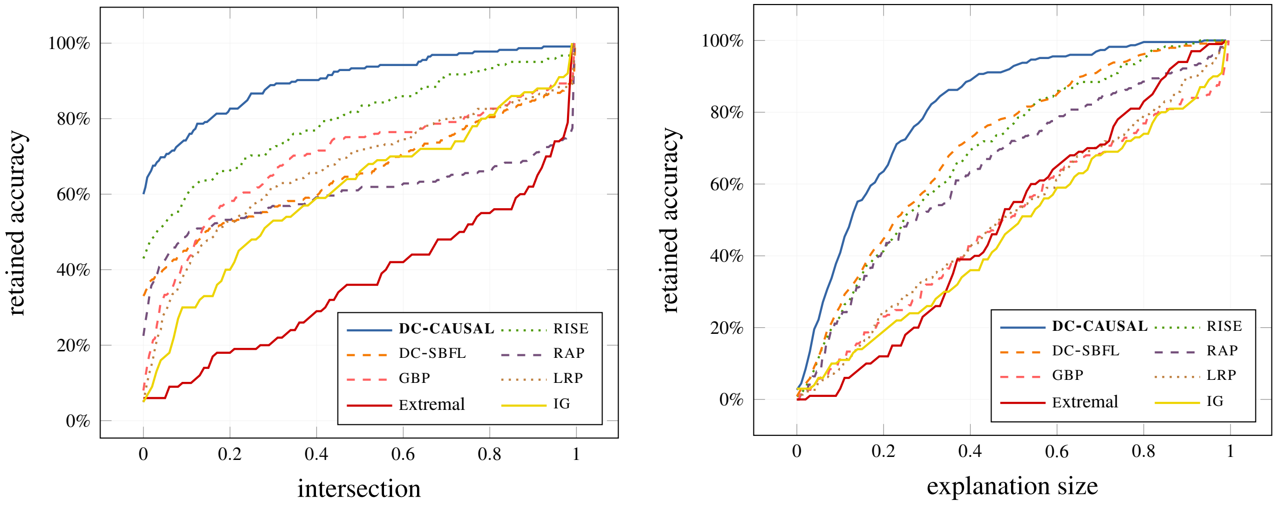
We configure DC-Causal to use four superpixels per partition. The termination conditions for the partition refinement in Alg. 2 are: 1) the height/width of a superpixel is smaller than $\frac{1}{10}$ of the input image's or 2) the four superpixels share the same responsibility. The parameter N of Alg. 3 is set to 50.

MS-COCO images with partial occlusions In this part of the evaluation, we consider explanations for classifications done by the compositional net [18] for partially occluded input images from the MS-COCO dataset [20]. Due to the lack of ground truth for this data set (Fig. 7), we use the (normalised) size of the explanation as proxy for quality. On average, the explanations from DC-Causal contain less than 13% of the total pixels and this compares favourably to 37.8% for DC-SBFL. DC-Causal's ranking yields 80% correct classification results when using the Compositional Net on only 20% of the pixels of the input image. This is more than a 20% improvement over DC-SBFL.

"Photo Bombing" images Similarly to the images with partial occlusion used in Section 5, we create an image set named "Photo Bombing". We plant occlusions (aka "photobombers") into ImageNet images and we record these occlusion pixels so that we can measure the intersection of explanations with the occluded pixels. Examples (and the corresponding DC-Causal explanations) from the Photo Bombing data set are given in Fig. 5.

Fig. 6 gives the results on the photobombing images. The y-axis is the retained accuracy of the classifier on the explanation part of the image (with the rest masked). We observe

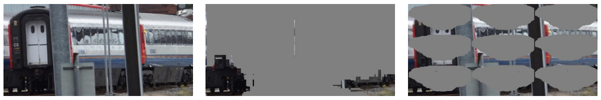
[3] The experimental data in this section and more results (e.g., for different configurations of the algorithm and explaining misclassifications) can be found at https://www.cprover.org/deepcover/.



(a)

(b)

Figure 6: Intersection between the explanation and the occlusion (smaller is better) and the size of the explanation (smaller is better) on the Photo Bombing dataset



(a) 'train'

(b) DC-Causal

(c) DC-SBFL

Figure 7: Explaining the partially occluded 'train' (a): DC-Causal (b) vs. DC-SBFL (c)

that more than 60% of the DC-Causal explanations do not have any overlap with the artificially planted occlusions, which is 20% better than than the second best tool RISE. The explanation size from DC-Causal is also consistently smaller than other tools, as shown in Fig. 6b. Interestingly, even though DC-Causal constructs smaller explanations than RISE, its explanations have a higher overlap with the occlusions, thus illustrating the necessity of using more than one quality measure for assessing the quality of explanations.

Masking colors When using DC-Causal (or any explanation algorithm that uses input perturbation), we need a masking color for removing parts of the input image. We argue empirically that the choice of masking color has little to no impact on the performance of our algorithm. To this end, we have run our experiments with a number of masking colors, ranging from black $(0,0,0)$ to white $(255,255,255)$. We measured the mean intersection between DC-Causal's explanations and the occlusions for each masking color. We have observed that the influence of the masking color is less than 4%. The full results for this experiment are on the project website.

"Roaming Panda" images To determine the performance of DC-Causal on images without occlusion, we have performed an experiment on the "Roaming Panda" dataset [30]. The classifier detects the Panda, which is superimposed and therefore never occluded in this dataset. We have compared DC-Causal with the other explanation tools by measuring the intersection of union (IoU) between the explanation and the panda part (the larger the better). DC-Causal achieved a (very close) third place, demonstrating that its performance on non-occluded images is comparable to the best explanation tools. The full results are available on the project website.

6. Conclusions

We propose a causal approach for explaining the output of image classifiers. Based on the definitions in [15], an explanation is a minimal subset of pixels of the image that is sufficient for the classification. As the exact computation is intractable [5], we describe a modular and parallelizable algorithm for computing an approximation to the explanation using a *causal ranking* of parts of the input image. Our experiments demonstrate that DC-Causal produces accurate results for partially occluded images, which pose a challenge to other explanation tools.

Acknowledgments

The authors acknowledge funding from the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1) and the UKRI Strategic Priorities Fund to the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (EP/V026607/1).

References

[1] 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), 2015. 2, 7

[2] Elie Bienenstock, Stuart Geman, and Daniel Potter. Compositionality, MDL priors, and object recognition. In *Advances in Neural Information Processing Systems*, pages 838–844, 1997. 2

[3] Urszula Chajewska and Joseph Y. Halpern. Defining explanation in probabilistic systems. In *Uncertainty in Artificial Intelligence (UAI)*, pages 62–71. Morgan Kaufmann, 1997. 1

[4] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning (ICML)*, volume 80, pages 882–891. PMLR, 2018. 2

[5] Hana Chockler and Joseph Y. Halpern. Responsibility and blame: A structural-model approach. *J. Artif. Intell. Res.*, 22:93–115, 2004. 3, 8

[6] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (S&P)*, pages 598–617. IEEE, 2016. 2

[7] Thomas Eiter and Thomas Lukasiewicz. Complexity results for structure-based causality. *Artif. Intell.*, 142(1):53–89, 2002. 3

[8] Thomas Eiter and Thomas Lukasiewicz. Complexity results for explanations in the structural-model approach. *Artif. Intell.*, 154(1-2):145–198, 2004. 4

[9] Sanja Fidler, Marko Boben, and Ales Leonardis. Learning a hierarchical compositional shape vocabulary for multi-class object representation. *arXiv*, 1408.5516, 2014. 2

[10] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *International Conference on Computer Vision (ICCV)*, pages 2950–2958. IEEE, 2019. 2, 7

[11] Peter Gärdenfors. *Knowledge in Flux*. MIT Press, 1988. 1

[12] David Gunning. Explainable artificial intelligence (XAI) – program information. https://www.darpa.mil/program/explainable-artificial-intelligence, 2017. Defense Advanced Research Projects Agency. 1

[13] Joseph Y. Halpern. A modification of the Halpern–Pearl definition of causality. In *Proceedings of IJCAI*, pages 3022–3033. AAAI Press, 2015. 3

[14] Joseph Y. Halpern and Judea Pearl. Causes and explanations: a structural-model approach. Part I: Causes. *British Journal for the Philosophy of Science*, 56(4), 2005. 3, 4

[15] Joseph Y. Halpern and Judea Pearl. Causes and explanations: a structural-model approach. Part II: Explanations. *British Journal for the Philosophy of Science*, 56(4), 2005. 4, 8

[16] Carl Gustav Hempel. *Aspects of Scientific Explanation*. Free Press, 1965. 1

[17] David Hume. *A Treatise of Human Nature*. John Noon, London, 1739. 3

[18] Adam Kortylewski, Ju He, Qing Liu, and Alan L Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8940–8949, 2020. 6, 7

[19] Adam Kortylewski, Qing Liu, Huiyu Wang, Zhishuai Zhang, and Alan Yuille. Combining compositional models and deep networks for robust object classification under occlusion. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1333–1341, 2020. 2, 7

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[21] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pages 4765–4774, 2017. 2, 6

[22] Woo-Jeoung Nam, Shir Gur, Jaesik Choi, Lior Wolf, and Seong-Whan Lee. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 2501–2508, 2020. 2, 7

[23] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988. 1

[24] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*. BMVA Press, 2018. 2, 7

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144. ACM, 2016. 2

[26] Wesley C. Salmon. *Four Decades of Scientific Explanation*. University of Minnesota Press, 1989. 1

[27] 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 *International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, 2017. 2

[28] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning (ICML)*, volume 70, pages 3145–3153. PMLR, 2017. 2

[29] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (Workshop Track)*, 2015. 2, 7

[30] Youcheng Sun, Hana Chockler, Xiaowei Huang, and Daniel Kroening. Explaining image classifiers using statistical fault localization. In *ECCV, Part XXVIII*, volume 12373 of *LNCS*, pages 391–406. Springer, 2020. 2, 6, 7, 8

[31] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 2, 7

[32] Jianyu Wang, Zhishuai Zhang, Cihang Xie, Vittal Premachandran, and Alan Yuille. Unsupervised learning of object semantic parts from internal states of CNNs by population encoding. *arXiv preprint arXiv:1511.06855*, 2015. 2, 7

[33] Mingqing Xiao, Adam Kortylewski, Ruihai Wu, Siyuan Qiao, Wei Shen, and Alan Yuille. TDAPNet: Prototype network with recurrent top-down attention for robust object classification under partial occlusion. *arXiv preprint arXiv:1909.03879*, 2019. 2

[34] Zhishuai Zhang, Cihang Xie, Jianyu Wang, Lingxi Xie, and Alan L. Yuille. DeepVoting: A robust and explainable deep network for semantic part detection under partial occlusion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1372–1380. IEEE, 2018. 2

[35] Hongru Zhu, Peng Tang, Jeongho Park, Soojin Park, and Alan Yuille. Robustness of object recognition under extreme occlusion in humans and computational models. *arXiv preprint arXiv:1905.04598*, 2019. 2