Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks

Jörg Wagner[1,2] Jan Mathias Köhler[1] Tobias Gindele[1,*] Leon Hetzel[1,*]
Jakob Thaddäus Wiedemer[1,*] Sven Behnke[2]
[1]Bosch Center for Artificial Intelligence (BCAI), Germany [2]University of Bonn, Germany
Joerg.Wagner3@de.bosch.com; behnke@cs.uni-bonn.de

Abstract

*To verify and validate networks, it is essential to gain insight into their decisions, limitations as well as possible shortcomings of training data. In this work, we propose a post-hoc, optimization based visual explanation method, which highlights the evidence in the input image for a specific prediction. Our approach is based on a novel technique to defend against adversarial evidence (*i.e. *faulty evidence due to artefacts) by filtering gradients during optimization. The defense does not depend on human-tuned parameters. It enables explanations which are both fine-grained and preserve the characteristics of images, such as edges and colors. The explanations are interpretable, suited for visualizing detailed evidence and can be tested as they are valid model inputs. We qualitatively and quantitatively evaluate our approach on a multitude of models and datasets.*

1. Introduction

Convolutional Neural Networks (CNNs) have proven to produce state-of-the-art results on a multitude of vision benchmarks, such as ImageNet [34], Caltech [12] or Cityscapes [9] which led to CNNs being used in numerous real-world systems (*e.g.* autonomous vehicles) and services (*e.g.* translation services). Though, the use of CNNs in safety-critical domains presents engineers with challenges resulting from their black-box character. A better understanding of the inner workings of a model provides hints for improving it, understanding failure cases and it may reveal shortcomings of the training data. Additionally, users generally trust a model more when they understand its decision process and are able to anticipate or verify outputs [30].

To overcome the interpretation and transparency disadvantage of black-box models, post-hoc explanation methods have been introduced [53, 35, 42, 49, 32, 17, 11]. These methods provide explanations for individual predictions and thus help to understand on which evidence a model bases its decisions. The most common form of explanations are visual, image-like representations, which depict the important pixels or image regions in a human interpretable manner.

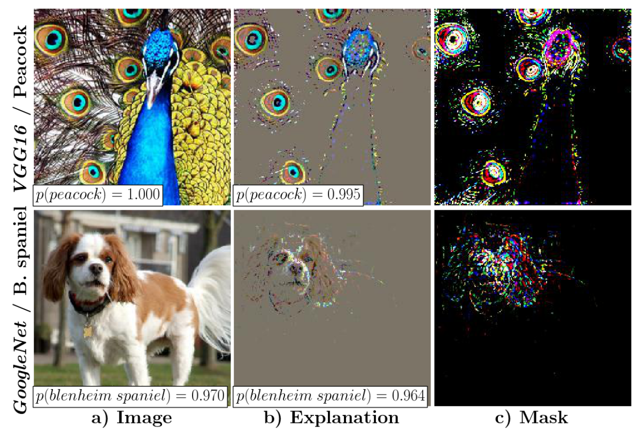

Figure 1: Fine-grained explanations computed by removing irrelevant pixels. a) Input image with softmax score $p(c_{ml})$ of the most-likely class. Our method tries to find a sparse mask (c) with irrelevant pixels set to zero. The resulting explanation (b), *i.e.*: 'image $\times$ mask', is optimized in the image space and, thus, can directly be used as model input. The parameter λ is optimized to produce an explanation with a softmax score comparable to the image.

In general, an explanation should be easily interpretable (Sec. 4.1). Additionally, a visual explanation should be class discriminative and fine-grained [35] (Sec. 4.2). The latter property is particularly important for classification tasks in the medical [20, 18] domain, where fine structures (*e.g.* capillary hemorrhages) have a major influence on the classification result (Sec. 5.2). Besides, the importance of different color channels should be captured, *e.g.* to uncover

[*]contributed while working at BCAI. We additionally thank Volker Fischer, Michael Herman, Anna Khoreva for discussions and feedback.

a color bias in the training data (Sec. 4.3).

Moreover, explanations should be faithful, meaning they accurately explain the function of the black-box model [35]. To evaluate the faithfulness (Sec. 5.1), recent work [35, 32, 7] introduce metrics which are based on model predictions of explanations. To be able to compute such metrics without having to rely on proxy measures [35], it is beneficial to employ explanation methods which directly generate valid model inputs (*e.g.* a perturbed version of the image).

A major concern of optimization based visual explanation methods is adversarial evidence, *i.e.* faulty evidence generated by artefacts introduced in the computation of the explanation. Therefore, additional constraints or regularizations are used to prevent such faulty evidence [17, 11, 14]. A drawback of these defenses are added hyperparameters and the necessity of either a reduced resolution of the explanation or a smoothed explanation (Sec. 3.2), thus, they are not well suited for displaying fine-grained evidence.

Our main contribution is a new adversarial defense technique which selectively filters gradients in the optimization which would lead to adversarial evidence otherwise (Sec. 3.2). Using this defense, we extend the work of [17] and propose a new fine-grained visual explanation method (FGVis). The proposed defense is not dependend on hyperparameters and is the key to produce fine-grained explanations (Fig. 1) as no smoothing or regularizations are necessary. Like other optimization-based approaches, FGVis computes a perturbed version of the original image, in which either all irrelevant or the most relevant pixels are removed. The resulting explanations (Fig 1 b) are valid model inputs and their faithfulness can, thus, be directly verified (as in methods from [17, 14, 6, 11]). Moreover, they are additionally fine-grained (as in methods from [35, 38, 48, 42]). To the best of our knowledge, this is the first method to be able to produce fine-grained explanations directly in the image space. We evaluate our defense (Sec. 3.2) and FGVis (Sec. 4 and 5) qualitatively and quantitatively.

2. Related Work

Various methods to create explanations have been introduced. Thang *et al.* [50] and DU *et al.* [13] provide a survey of these. In this section, we give an overview of explanation methods which generate visual, image-like explanations.

Backpropagation Based Methods (BBM). These methods generate an importance measure for each pixel by backpropagating an error signal to the image. Simonyan *et al.* [38], which build on work of Baehrens *et al.* [5], use the derivative of a class score with respect to the image as an importance measure. Similar methods have been introduced in Zeiler *et al.* [48] and Springenberg *et al.* [42], which additionally manipulate the gradient when backpropagating through ReLU nonlinearities. Integrated Gradients [43] additionally accumulates gradients along a path from a base image to the input image. SmoothGrad [40] and VarGrad [1] visually sharpen explanations by combining multiple explanations of noisy copies of the image. Other BBMs such as Layer-wise Relevance Propagation [4], DeepLift [37] or Excitation Backprop [49] utilize top-down relevancy propagation rules. BBMs are usually fast to compute and produce fine-grained importance/relevancy maps. However, these maps are generally of low quality [11, 14] and are less interpretable. To verify their faithfulness it is necessary to apply proxy measures or use pre-processing steps, which may falsify the result.

Activation Based Methods (ABM). These approaches use a linear combination of activations from convolutional layers to form an explanation. Prominent methods of this category are CAM (Class Activation Mapping) [53] and its generalizations Grad-CAM [35] and Grad-CAM++ [7]. These methods mainly differ in how they calculate the weights of the linear combination and what restrictions they impose on the CNN. Extensions of such approaches have been proposed in Selvaraju *et al.* [35] and Du *et al.* [14], which combine ABMs with backpropagation or perturbation based approaches. ABMs generate easy to interpret heat-maps which can be overlaid on the image. However, they are generally not well suited to visualize fine-grained evidence or color dependencies. Additionally, it is not guaranteed that the resulting explanations are faithful and reflect the decision making process of the model [14, 35].

Perturbation Based Methods (PBM). Such approaches perturb the input and monitor the prediction of the model. Zeiler *et al.* [48] slide a grey square over the image and use the change in class probability as a measure of importance. Several approaches are based on this idea, but use other importance measures or occlusion strategies. Petsiuk *et al.* [32] use randomly sampled occlusion masks and define importance based on the expected model score over masks. LIME [33] uses a super-pixel based occlusion strategy and a surrogate model to compute importance scores. Further super-pixel or segment based methods are introduced in Seo *et al.* [36] and Zhou *et al.* [52]. The so far mentioned approaches do not need access to the internal state or structure of the model. Though, they are often quite time consuming and only generate coarse explanations.

Other PBMs generate an explanation by optimizing for a perturbed version of the image [11, 17, 14, 6]. The perturbed image $\mathbf{e}$ is defined by $\mathbf{e} = \mathbf{m} \cdot \mathbf{x} + (1 - \mathbf{m}) \cdot \mathbf{r}$, where $\mathbf{m}$ is a mask, $\mathbf{x}$ the input image, and $\mathbf{r}$ a reference image containing little information (Sec. 3.1). To avoid adversarial evidence, these approaches need additional regularizations [17], constrain the explanation (*e.g.* optimize for a coarse mask [6, 17, 14]), introduce stochasticity [17], or utilize regularizing surrogate models [11]. These approaches generate easy to interpret explanations in the image space, which are valid model inputs and faithful (*i.e.* a faithfulness measure is incorporated in the optimization).

Our method also optimizes for a perturbed version of the input. Compared to existing approaches we propose a new adversarial defense technique which filters gradients during optimization. This defense does not need hyperparameters which have to be fine-tuned. Besides, we optimize each pixel individually, thus, the resulting explanations have no limitations on the resolution and are fine-grained.

3. Explaining Model Predictions

Explanations provide insights into the decision-making process of a model. The most universal form of explanations are *global* ones which characterize the overall model behavior. *Global* explanations specify for all possible model inputs the corresponding output in an intuitive manner. A decision boundary plot of a classifier in a low-dimensional vector space, for example, represents a *global* explanation. For high-dimensional data and complex models, it is practically impossible to generate such explanations. Current approaches therefore utilize *local* explanations[1], which focus on individual inputs. Given one data point, these methods highlight the evidence on which a model bases its decisions. As outlined in Sec. 2, the definition of highlighting depends on the used explanation method. In this work, we follow the paradigm introduced in [17] and directly optimize for a perturbed version of the input image. Such an approach has several advantages: 1) The resulting explanations are interpretable due to their image-like nature; 2) Explanations represent valid model inputs and are thus testable; 3) Explanations are optimized to be faithful. In Sec. 3.1 we briefly review the general paradigm of optimization based explanation methods before we introduce our novel adversarial defense technique in Sec. 3.2.

3.1. Perturbation based Visual Explanations

Following the paradigm of optimization based explanation methods, which compute a perturbed version of the image [17, 14, 6, 11], an explanation can be defined as:

Explanation by Preservation: The smallest region of the image which must be retained to preserve the original model output (*i.e.* minimal sufficient evidence).

Explanation by Deletion: The smallest region of the image which must be deleted to change the model output.

To formally derive an explanation method based on this paradigm, we assume that a CNN f_{cnn} is given which maps an input image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ to an output $\mathbf{y}_x = f_{cnn}(\mathbf{x}; \theta_{cnn})$. The ouput $\mathbf{y}_x \in \mathbb{R}^C$ is a vector representing the softmax scores y_x^c of the different classes c. Given an input image $\mathbf{x}$, an explanation $\mathbf{e}^*_{c_T}$ of a target class c_T (*e.g.* the most-likely class $c_T = c_{ml}$) is computed by removing either relevant (*deletion*) or irrelevant, not supporting c_T, information (*preservation*) from the image. Since it is not possible to remove information without replacing it, and we do not have access to the image generating process, we have to use an approximate removal operator [17]. A common approach is to use a mask based operator Φ, which computes a weighted average between the image $\mathbf{x}$ and a reference image $\mathbf{r}$, using a mask $\mathbf{m}_{c_T} \in [0,1]^{3\times H\times W}$:

$$\mathbf{e}_{c_T} = \Phi(\mathbf{x}, \mathbf{m}_{c_T}) = \mathbf{x}\cdot\mathbf{m}_{c_T} + (1-\mathbf{m}_{c_T})\cdot\mathbf{r}. \quad (1)$$

Common choices for the reference image are constant values (*e.g.* zero), a blurred version of the original image, Gaussian noise, or sampled references of a generative model [17, 14, 6, 11]. In this work, we take a zero image as reference. In our opinion, this reference produces the most pleasing visual explanations, since irrelevant image areas are set to zero[2] (Fig. 1) and not replaced by other structures. In addition, the zero image (and random image) carry comparatively little information and lead to a model prediction with a high entropy. Other references, such as a blurred version of the image, usually result in lower prediction entropies, as shown in Sec. A3.1. Due to the additional computational effort, we have not considered model-based references as proposed in Chang *et al.* [6].

In addition, a similarity metric $\varphi(y_x^{c_T}, y_e^{c_T})$ is needed, which measures the consistency of the model output generated by the explanation $y_e^{c_T}$ and the output of the image $y_x^{c_T}$ with respect to a target class c_T. This similarity metric should be small if the explanation preserves the output of the target class and large if the explanation manages to significantly drop the probability of the target class [17]. Typical choices for the metric are the cross-entropy with the class c_T as a hard target [24] or the negative softmax score of the target class c_T. The similarity metric ensures that the explanation remains faithful to the model and thus accurately explains the function of the model, this property is a major advantage of PBMs.

Using the mask based definition of an explanation with a zero image as reference ($\mathbf{r} = \mathbf{0}$) as well as the similarity metric, a *preserving explanation* can be computed by:

$$\begin{aligned} \mathbf{e}^*_{c_T} &= \mathbf{m}^*_{c_T}\cdot\mathbf{x}, \\ \mathbf{m}^*_{c_T} &= \underset{\mathbf{m}_{c_T}}{\arg\min}\{\varphi(y_x^{c_T}, y_e^{c_T}) + \lambda\cdot\|\mathbf{m}_{c_T}\|_1\}. \end{aligned} \quad (2)$$

We will refer to the optimization in Eq. 2 as the *preservation game*. Masks (Fig. 2/b2)[3] generated by this game are sparse (*i.e.* many pixels are zero / appear black; enforced by minimizing $\|\mathbf{m}_{c_T}\|_1$) and only contain large values at most important pixels. The corresponding explanation is computed by multiplying the mask with the image (Fig. 2/c2).

[1] For the sake of brevity, we will use the term explanations as a synonym for *local* explanations throughout this work.

[2] Tensors $\mathbf{x}, \mathbf{e}, \mathbf{r}$ are assumed to be normalized according to the training of the CNN. A value of zero for these thus corresponds to a grey color (*i.e.* the color of the data mean).

[3] Fig. 2/b2: Figure 2, column b, 2nd row

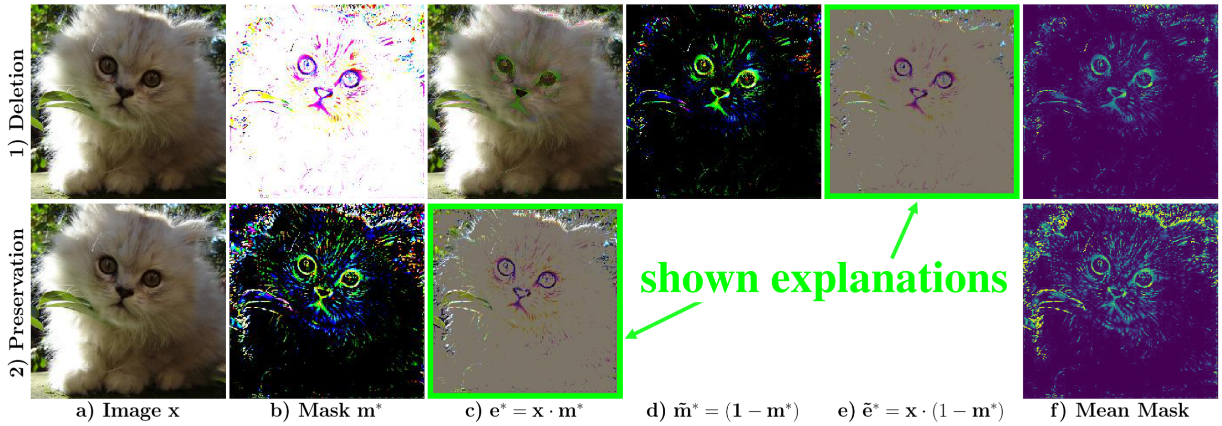

Figure 2: Visualization types calculated for *VGG* using *deletion / preservation game*. For the *repression / generation game* the same characteristics hold. Subscript c_T ommited to ease readability. a) Input image. b) Mask obtained by the optimization. Colors in a *deletion* mask are complementary to the image colors. c) Explanation directly obtained by the optimization. d) Complementary mask with a true-color representation for the *deletion game*. e) Explanation highlighting the important evidence for the *deletion game*. f) Mean mask: mask / comp. mask averaged over colors. — To underline important evidence, we use $\mathbf{e}$ for the explanation of the *preservation / generation game* and $\tilde{\mathbf{e}}$ for the *deletion / repression game*.

Alternatively, we can compute a *deleting explanation* using:

$$\begin{aligned} \mathbf{e}^*_{c_T} &= \mathbf{m}^*_{c_T} \cdot \mathbf{x}, \\ \mathbf{m}^*_{c_T} &= \arg\max_{\mathbf{m}_{c_T}} \{\varphi(y^{c_T}_x, y^{c_T}_e) + \lambda \cdot \|\mathbf{m}_{c_T}\|_1\}. \end{aligned} \tag{3}$$

This optimization will be called *deletion game* henceforward. Masks (Fig. 2 / b1) generated by this game contain mainly ones (*i.e.* appear white; enforced by maximizing $\|\mathbf{m}_{c_T}\|_1$ in Eq. 3) and only small entries at pixels, which provide the most prominent evidence for the target class. The colors in a mask of the *deletion game* are complementary to the image colors. To obtain a true-color representation analogous to the *preservation game*, one can alternatively visualize the complementary mask (Fig. 2 / d1): $\tilde{\mathbf{m}}^*_{c_T} = (\mathbf{1} - \mathbf{m}^*_{c_T})$. A resulting explanation of the *deletion game*, as defined in Eq. 3, is visualized in Fig. 2 / c1. This explanation is visually very similar to the original image as only a few pixels need to be deleted to change the model output. In the remaining of the paper for better visualization, we depict a modified version of the explanation for the *deletion game*: $\tilde{\mathbf{e}}^*_{c_T} = \mathbf{x} \cdot (\mathbf{1} - \mathbf{m}^*_{c_T})$. This explanation has the same properties as the one of the *preservation game*, *i.e.* it only highlights the important evidence. We observe that the *deletion game* generally produces sparser explanations compared to the *preservation game*, as less pixels have to be removed to delete evidence for a class than to maintain evidence by preserving pixels.

To solve the optimization in Eq. 2 and Eq. 3, we utilize Stochastic Gradient Descent and start with an explanation $\mathbf{e}^0_{c_T} = \mathbf{1} \cdot \mathbf{x}$ identical to the original image (*i.e.* a mask initialized with ones). As an alternative initialization of the masks, we additionally explore a zero initialization $\mathbf{m}^0_{c_T} = \mathbf{0}$. In this setting the initial explanation contains no evidence towards any class and the optimization iteratively has to add relevant (*generation game*) or irrelevant, not supporting the class c_T, information (*repression game*). The visualizations of the *generation game* are equivalent to those of the *preservation game*, the same holds for the *deletion* and *repression game*. In our experiments the *deletion game* produces the most fine-grained and visually pleasing explanations. Compared to the other games it usually needs the least amount of optimization iterations since we start with $\mathbf{m}^0_{c_T} = \mathbf{1}$ and comparatively few mask values have to be changed to delete the evidence for the target class. A comparison and additional characteristics of the four optimization settings (*i.e.* games) are included in Sec. A3.5.

3.2. Defending against Adversarial Evidence

CNNs have been proven susceptible to adversarial images [45, 19, 27], i.e. a perturbed version of a correctly classified image crafted to fool a CNN. Due to the computational similarity of adversarial methods and optimization based visual explanation approaches, adversarial noise is also a concern for the latter methods and one has to ensure that an explanation is based on true evidence present in the image and not on false adversarial evidence introduced during optimization. This is particularly true for the *generation/repression game* as their optimization start with $\mathbf{m}^0_{c_T} = \mathbf{0}$ and iteratively adds information.

[17] and [11] showed the vulnerability of optimization based explanation methods to adversarial noise. To avoid adversarial evidence, explanation methods use stochastic operations [17], additional regularizations [17, 11], optimize on a low-resolution mask with upsampling of the computed mask [17, 14, 6], or utilize a regularizing surrogate

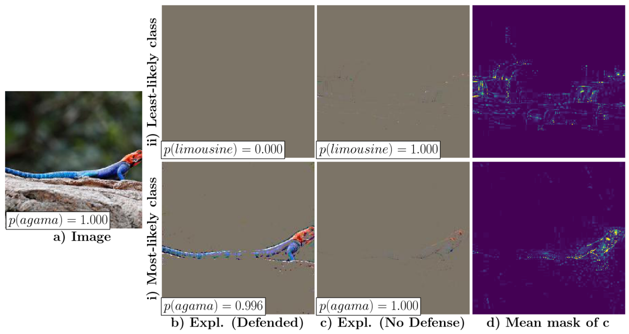

Figure 3: Explanations computed for the adversarial class *limousine* and the predicted class *agama* using the *generation game* and *VGG16* with and without our adversarial defense. An adversarial for class *limousine* can only be computed without the defense. d) Mean mask enhanced by a factor of 7 to show small adversarial structures.

model [11]. In general, these operations impede the generation of adversarial noise by obscuring the gradient direction in which the model is susceptible to false evidence, or by constraining the search space for potential adversarials. These techniques help to reduce adversarial evidence, but also introduce new drawbacks: 1) Defense capabilities usually depend on human-tuned parameters; 2) Explanations are limited to being low resolution and/or smooth, which prevents fine-grained evidence from being visualized.

A novel Adversarial Defense. To overcome these drawbacks, we propose a novel adversarial defense which filters gradients during backpropagation in a targeted way. The basic idea of our approach is: A neuron within a CNN is only allowed to be activated by the explanation $\mathbf{e}_{c_T}$ if the same neuron was also activated by the original image $\mathbf{x}$. If we regard neurons as indicators for the existence of features (*e.g*. edges, object parts, ...), the proposed constraint enforces that the explanation $\mathbf{e}_{c_T}$ can only contain features which exist at the same location in the original image $\mathbf{x}$. By ensuring that the allowed features in $\mathbf{e}_{c_T}$ are a subset of the features in $\mathbf{x}$ it prevents the generation of new evidence.

This defense technique can be integrated in the introduced explanation methods via an optimization constraint:

$$\begin{cases} 0 \leq h_i^l(\mathbf{e}_{c_T}) \leq h_i^l(\mathbf{x}), & \text{if } h_i^l(\mathbf{x}) \geq 0, \\ 0 \geq h_i^l(\mathbf{e}_{c_T}) \geq h_i^l(\mathbf{x}), & \text{otherwise,} \end{cases} \tag{4}$$

where h_i^l is the activation of the i-th neuron in the l-th layer of the network after the nonlinearity. For brevity, the index i references one specific feature at one spatial position in the activation map. This constraint is applied after all nonlinearity-layers (*e.g*. ReLU-Layers) of the network, besides the final classification layer. It ensures that the absolute value of activations can only be reduced towards values representing lower information content (we assume that zero activations have the lowest information as commonly applied in network pruning [22]). To solve the optimization with subject to Eq. 4, one could incorporate the constraints via a penalty function in the optimization loss. The drawback is one additional hyperparameter. Alternatively, one could add an additional layer $\bar{h}_i^l$ after each nonlinearity which ensures the validity of Eq. 4:

$$\begin{aligned} \bar{h}_i^l(\mathbf{e}_{c_T}) &= \min(bu, \max(bl, h_i^l(\mathbf{e}_{c_T}))), \\ bu &= \max(0, h_i^l(\mathbf{x})), \\ bl &= \min(0, h_i^l(\mathbf{x})), \end{aligned} \tag{5}$$

where $h_i^l(\mathbf{e}_{c_T})$ is the actual activation of the original nonlinearity-layer and $\bar{h}_i^l(\mathbf{e}_{c_T})$ the adjusted activation after ensuring the bounds bu, bl of the original input. For instance, for a ReLU nonlinearity, the upper bound bu is equal to $h_i^l(\mathbf{x})$ and the lower bound bl is zero. We are not applying this method as it changes the architecture of the model which we try to explain. Instead, we clip gradients in the backward pass of the optimization, which lead to a violation of Eq. 4. This is equivalent to adding an additional clipping-layer after each nonlinearity which acts as the identity in the forward pass and uses the gradient update of Eq. 5 in the backward pass. When backpropagating an error-signal $\bar{\gamma}_i^l$ through the clipping-layer, the gradient update rule for the resulting error γ_i^l is defined by:

$$\gamma_i^l = \bar{\gamma}_i^l \cdot [h_i^l(\mathbf{e}_{c_T}) \leq bu] \cdot [h_i^l(\mathbf{e}_{c_T}) \geq bl], \tag{6}$$

where $[\,\cdot\,]$ is the indicator function and bl, bu the bounds computed in Eq. 5. This clipping only affects the gradients of the similarity metric $\varphi(\cdot\,,\cdot)$ which are propagated through the network. The proposed gradient clipping does not add hyperparameters and keeps the original structure of the model during the forward pass. Compared to other adversarial defense techniques ([11], [17], [6]), it imposes no constraint on the explanation (*e.g*. resolution/smoothness constraints), enabling fine-grained explanations.

Validating the Adversarial Defense. To evaluate the performance of our defense, we compute an explanation for a class c_A for which there is no evidence in the image (*i.e*. it is visually not present). We approximate c_A with the least-likely class c_{ll} considering only images which yield very high predictive confidence for the true class $p(c_{true}) \geq 0.995$. Using c_{ll} as the target class, the resulting explanation method without defense is similar to an adversarial attack (the *Iterative Least-Likely Class Method* [27]).

A correct explanation for the adversarial class c_A should be "empty" (*i.e*. grey), as seen in Fig. 3 b, top row, when using our adversarial defense. If, on the other hand, the explanation method is susceptible to adversarial noise, the optimization procedure should be able to perfectly generate an explanation for any class. This behavior can be seen in Fig. 3 c, top row. The shown explanation for the adversarial

Model	No Defense	Defended
VGG16 [39]	100.0 %	0.2 %
AlexNet [26]	100.0 %	0.0 %
ResNet50 [23]	100.0 %	0.0 %
GoogleNet [44]	100.0 %	0.0 %

Table 1: Ratio how often an adversarial class c_A was generated, using the *generation game* with no sparsity loss on *VGG16* with and without our defense.

class (c_A: *limousine*) contains primarily artificial structures and is classified with a probability of 1 as *limousine*.

We also depict the explanation of the predicted class (c_{pred}: *agama*). The explanation with our defense results in a meaningful representation of the *agama* (Fig. 3 b, bottom row); without defense (Fig. 3 c / d, bottom row) it is much more sparse. As there is no constraint to change pixel values arbitrarily, we assume the algorithm introduces additional structures to produce a sparse explanation.

A quantitative evaluation of the proposed defense is reported in Tab. 1. We generate explanations for 1000 random ImageNet validation images and use a class c_A as the explanation target[4]. To ease the generation of adversarial examples, we set the sparsity loss to zero and only use the similarity metric which tries to maximize the probability of the target class c_A. Without an employed defense technique, the optimization is able to generate an adversarial explanation for 100% of the images. Applying our defense (Eq. 6), the optimization nearly never was able to do so. The two adverarial examples generated in *VGG16* have a low confidence, so we assume that there has been some evidence for the chosen class c_A in the image. Our proposed technique is thus well suited to defend against adversarial evidence.

4. Qualitative Results

Implementation details are stated in Sec. A2.

4.1. Interpretability

Comparison of methods. Using the *deletion game* we compute mean explanation masks for *GoogleNet* and compare these in Fig. 5 with state-of-the-art methods. Our method delivers the most fine-grained explanation by deleting important pixels of the target object. Especially explanations b), f), and g) are coarser and, therefore, tend to include background information not necessary to be deleted to change the original prediction. The majority of pixels highlighted by FGVis form edges of the object. This cannot be seen in other methods. The explanations from c) and d) are most similar to ours. However, our masks are computed to directly produce explanations which are viable network inputs and are, therefore, verifiable — The deletion of the highlighted pixels prevents the model from correctly predicting the object. This statement does not necessarily hold for explanations calculated with methods c) and d).

Architectural insights. As first noted in [31] explanations using backpropagation based approaches show a grid-like pattern for ResNet. In general, [31] demonstrate that the network structure influences the visualization and assume that for ResNet the skip connections play an important role in their explanation behavior. As shown in Fig 6 this pattern is also visible in our explanations to an even finer degree. Interestingly, the grid pattern is also visible to a lesser extent outside the object. A detailed investigation of this phenomenon is left for future research. See A3.4 for a comparison of explanations between models.

4.2. Class Discriminative / Fine-Grained

Visual explanation methods should be able to produce class discriminative (*i.e.* focus on one object) and fine-grained explanations [35]. To test FGVis with respect to these properties, we generate explanations for images containing two objects. The objects are chosen from highly different categories to ensure little overlapping evidence. In Fig. 4, we visualize explanations of three such images, computed using the *deletion game* and *GoogleNet*. Additional results can be found in Sec. A3.2.

FGVis is able to generate class discriminative explanations and only highlights pixels of the chosen target class. Even partially overlapping objects, as the elkhound and ball in Fig. 4, first row, or the bridge and schooner in Fig. 4,

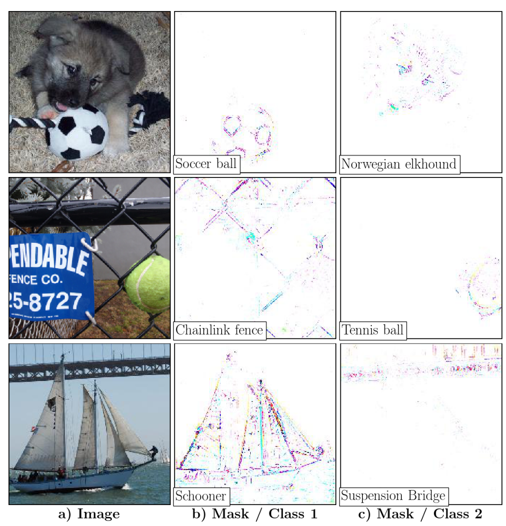

Figure 4: Explanation masks for images with multiple objects computed using the *deletion game* and *GoogleNet*. FGVis produces class discriminating explanations, even when objects partially overlap. Additionally, FGVis is able to visualize fine-grained details down to the pixel level.

[4]For c_A we used the least-likely class, as described before. We use the second least-likely class, if the least-likely class coincidentally matches the predicted class for the zero image.

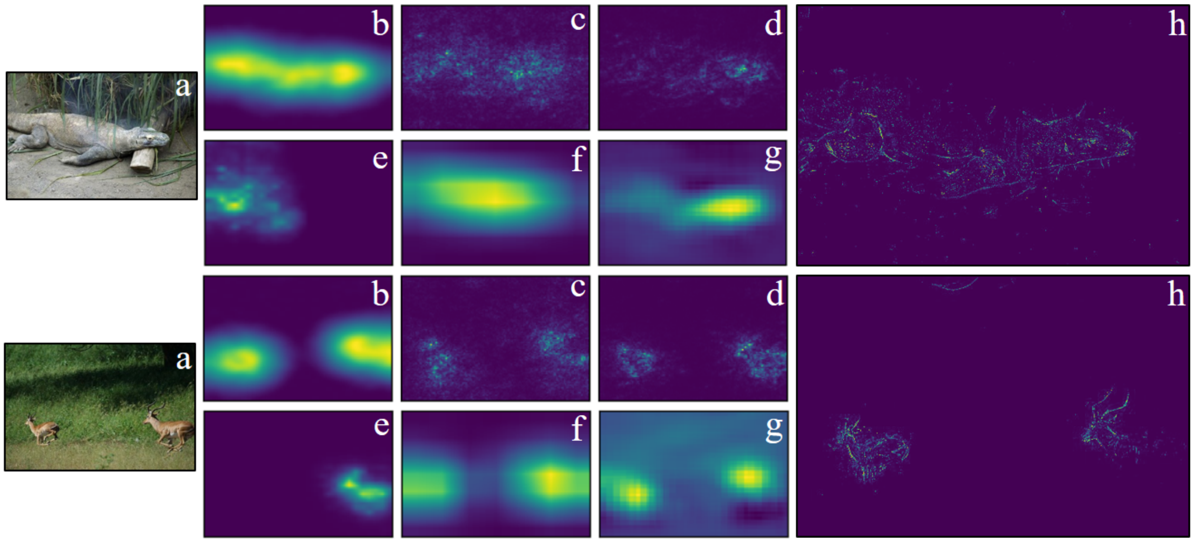

Figure 5: Comparison of mean explanation masks: a) Image, b) BBMP [17], c) Gradient [38], d) Guided Backprop [42] , e) Contrastive Excitation Backprop [49], f) Grad-CAM [35], g) Occlusion [48], h) FGVis (ours). The masks of all reference methods are based on work by [17]. Due to our detailed and sparse masks, we plot them in a larger size.

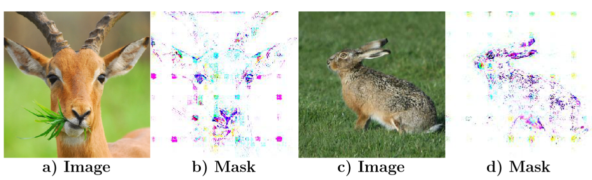

Figure 6: Visual explanations computed using the *deletion game* for *ResNet50*. The masks (b, d) show a grid-like pattern, as also observed in [31] for *ResNet50*.

third row, are correctly discriminated. One major advantage of FGVis is its ability to visualize fine-grained details. This property is especially visible in Fig 4, second row, which shows an explanation for the target class fence. Despite the fine structure of the fence, FGVis is able to compute a precise explanation which mainly contains fence pixels.

4.3. Investigating Biases of Training Data

An application of explanation methods is to identify a bias in the training data. Especially for safety-critical, high-risk domains (*e.g.* autonomous driving), such a bias can lead to failures if the model does not generalize to the real world.

Learned objects. One common bias is the coexistence of objects in images which can be depicted using FGVis. In Sec. A3.3, we describe such a bias in ImageNet for sports equipment appearing in combination with players.

Learned color. Objects are often biased towards specific colors. FGVis can give a first visual indication for the importance of different color channels. We investigate if a *VGG16* model trained on ImageNet shows such a bias using the *preservation game*. We focus on images of school buses and minivans and compare explanations (Fig. 7; all correctly predicted images in Fig. A6 and A8). Explanations of minivans focus on edges, not consistently preserving the color compared to school buses with yellow dominating those explanations. This is a first indication for the importance of color for the prediction of school buses.

To verify the qualitative finding, we quantitatively give an estimation of the color bias. As an evaluation we swap each of the three color channels *BGR* to either *RBG* or *GRB* and calculate the ratio of maintained true classifications on the validation data after the swap. For minivans 83.3% (averaged over *RBG* and *GRB*) of the 21 correctly classified images keep their class label, for school buses it is only 8.3% of 42 images. For 80 ImageNet classes at least 75% of images are no longer truly classified after the color swap. We show the results for the most and least affected 19 classes and minivan / school bus in Tab. A3.
To the best of our knowledge, FGVis is the first method used to highlight color channel importance.

5. Quantitative Results

5.1. Faithfulness of Explanations

The faithfulness of generated visual explanations to the underlying neural network is an important property of explanation methods [35]. To quantitatively compare the faithfulness of methods, Petsiuk *et al.* [32] proposed causal metrics which do not depend on human labels. These metrics are not biased towards human perception and are thus well suited to verify if an explanation correctly represents the evidence on which a model bases its prediction.

We use the deletion metric [32] to evaluate the faith-

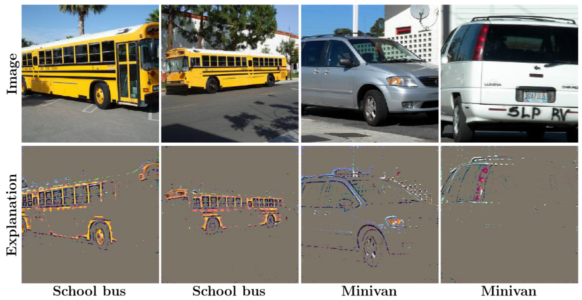

Figure 7: Explanations computed using the *preservation game* for *VGG16*. Explanations of the class minivan focus on edges, hardly preserving the color, compared to the class school bus, with yellow dominating the explanations.

fulnes of explanations generated by our method. This metric measures how the removal of evidence effects the prediction of the used model. The metric assumes that an importance map is given, which ranks all image pixels with respect to their evidence for the predicted class c_{ml}. By iteratively removing important pixels from the input image and measuring the resulting probability of the class c_{ml} a deletion curve can be generated, whose *area under the curve* AUC is used as a measure of faithfulness (Sec. A4.1).

In Tab. 2, we report the deletion metric of FGVis, computed on the validation split of ImageNet using different models. We use the *deletion game* to generate masks $\mathbf{m}_{ml}$, which determine the importance of each pixel. A detailed description of the experiment settings as well as additional figures, can be found in Sec. A4.1. FGVis outperforms the other explanation methods on both models by a large margin. This performance increase can be attributed to the ability of FGVis to visualize fine-grained evidence. All other approaches are limited to coarse explanations, either due to computational constraints or due to the used measures to avoid adversarial evidence. The difference between the two model architectures can most likely be attributed to the superior performance of *ResNet50*, resulting in on average higher softmax scores over all validation images.

Method	*ResNet50*	*VGG16*
Grad-Cam [35]	0.1232	0.1087
Sliding Window [48]	0.1421	0.1158
LIME[33]	0.1217	0.1014
RISE [32]	0.1076	0.0980
FGVis (ours)	**0.0644**	**0.0636**

Table 2: Deletion metric computed on the ImageNet validation dataset (lower is better). The results for all reference methods were taken from Petsiuk *et al*. [32].

5.2. Visual explanation for medical images

We evaluate FGVis on a real-world use case to identify regions in eye fundus images which lead a CNN to classify the image as being affected with referable diabetic retinopathy (RDR). Using the *deletion game* we derive a weakly-supervised approach to detect RDR lesions. The setup, used network, as well as details on the disease and training data are described in A4.2. To evaluate FGVis, the DiaretDB1 dataset [25] is used containing 89 fundus images with different lesion types, ground truth marked by four experts. To quantitatively judge the performance, we compare in Tab. 3 the image level sensitivity of detecting if a certain lesion type is present in an image. The methods [54, 28, 21, 29] use supervised approaches on image level without reporting a localization. [51] propose an unsupervised approach to extract salient regions. [18] use a comparable setting to ours applying CAM [53] in a weakly-supervised way to highlight important regions. To decide if a lesion is detected, [18] suggest an overlap of 50% between proposed regions and ground truth. As our explanation masks are fine-grained and the ground truth is coarse, we compare using a 25% overlap and for completeness report a 50% overlap.

It is remarkable that FGVis performs comparable or outperforms fully supervised approaches which are designed to detect the presence of one lesion type. The strength of FGVis is especially visible in detecting RSD, as these small lesions only cover some pixels in the image. In Fig. A21 we show fundus images, ground truth and our predictions.

Method	H	HE	SE	RSD
Zhou *et al.*[54]	94.4	-	-	
Liu *et al.*[28]	-	83.0	83.0	-
Haloi *et al.*[21]		**96.5**	-	-
Mane *et al.*[29]	-	-	-	**96.4**
Zhao *et al.* [51]	98.1	-	-	
Gondal *et al.*[18]	97.2	93.3	81.8	50
Ours (25% Overlap)	**100**	94.7	**90.0**	88.4
Ours (50% Overlap)	90.5	81.6	80.0	86.0

Table 3: Image level sensitivity in % (higher is better) for four different lesions H, HE, SE, RSD: Hemorrhages, Hard Exudates, Soft Exudates and Red Small Dots.

6. Conclusion

We propose a method which generates fine-grained visual explanations in the image space using on a novel technique to defend adversarial evidence. Our defense does not introduce hyperparameters. We show the effectivity of the defense on different models, compare our explanations to other methods, and quantitatively evaluate the faithfulness. Moreover, we underline the strength in producing class discriminative visualizations and point to characteristics in explanations of a *ResNet50*. Due to the fine-grained nature of our explanations, we achieve remarkable results on a medical dataset. Besides, we show the usability of our approach to visually indicate a color bias in training data.

References

[1] Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. In *Workshop at the International Conference on Learning Representations (ICLR)*, 2018. 2

[2] Ankita Agrawal, Charul Bhatnagar, and Anand Singh Jalal. A survey on automated microaneurysm detection in diabetic retinopathy retinal images. In *International Conference on Information Systems and Computer Networks (ISCON)*, pages 24–29. IEEE, 2013. 10

[3] R. Arunkumar and P. Karthigaikumar. Multi-retinal disease classification by reduced deep learning features. *Neural Computing and Applications*, 28(2):329–334, 2017. 10

[4] 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. 2

[5] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010. 2

[6] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. *arXiv e-prints*, page arXiv:1807.08024, Jul 2018. 2, 3, 4, 5

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 2

[8] E. Colas, A. Besse, A. Orgogozo, B. Schmauch, N. Meric, and E. Besse. Deep learning approach for diabetic retinopathy screening. *Acta Ophthalmologica*, 94(S256), 2016. 10

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 1

[10] Jorge Cuadros and George Bresnick. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of Diabetes Science and Technology*, 3(3):509–516, 2009. 10

[11] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6967–6976, 2017. 1, 2, 3, 4, 5

[12] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1

[13] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *arXiv e-prints*, page arXiv:1808.00033, Jul 2018. 2

[14] Mengnan Du, Ninghao Liu, Qingquan Song, and Xia Hu. Towards explanation of dnn-based prediction with guided feature inversion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1358–1367, 2018. 2, 3, 4

[15] EyePACS. https://www.kaggle.com/c/diabetic-retinopathy-detection. *assessed on 2018-09-23*, 2015. 10

[16] EyePACS. https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15617. assessed on 2018-09-23. 10

[17] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3429–3437, 2017. 1, 2, 3, 4, 5, 7

[18] Waleed M. Gondal, Jan M. Köhler, René Grzeszick, Gernot A. Fink, and Michael Hirsch. Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. In *IEEE International Conference on Image Processing (ICIP)*, pages 2069–2073, 2017. 1, 8, 10

[19] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 4

[20] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Journal of the American Medical Association (JAMA)*, 316(22):2402–2410, 2016. 1, 10

[21] Mrinal Haloi, Samarendra Dandapat, and Rohit Sinha. A gaussian scale space approach for exudates detection, classification and severity prediction. *arXiv e-prints*, page arXiv:1505.00737, May 2015. 8

[22] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1135–1143, 2015. 5

[23] 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 (CVPR)*, pages 770–778, 2016. 6

[24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv e-prints*, page arXiv:1503.02531, Mar 2015. 3

[25] Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iiris Sorri, Asta Raninen, Raija Voutilainen, Hannu Uusitalo, Heikki Kälviäinen, and Juhani Pietilä. The DIARETDB1 diabetic retinopathy database and evaluation protocol. In *British Machine Vision Conference (BMVC)*, pages 1–10, 2007. 8, 10

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012. 6, 1

[27] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv e-prints*, page arXiv:1607.02533, Jul 2016. 4, 5

[28] Qing Liu, Beiji Zou, Jie Chen, Wei Ke, Kejuan Yue, Zailiang Chen, and Guoying Zhao. A location-to-segmentation strategy for automatic exudate segmentation in colour retinal fundus images. *Computerized Medical Imaging and Graphics*, 55:78–86, 2017. 8

[29] Vijay M Mane, Ramish B Kawadiwale, and DV Jadhav. Detection of red lesions in diabetic retinopathy affected fundus images. In *IEEE International Advance Computing Conference (IACC)*, pages 56–60, 2015. 8

[30] Rowan McAllister, Yarin Gal, Alex Kendall, Mark Van Der Wilk, Amar Shah, Roberto Cipolla, and Adrian Vivian Weller. Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2017. 1

[31] Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. *arXiv e-prints*, page arXiv:1805.07039, May 2018. 6, 7

[32] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018. 1, 2, 7, 8, 10

[33] 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. 2, 8

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1

[35] 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 (ICCV)*, pages 618–626, 2017. 1, 2, 6, 7, 8

[36] Dasom Seo, Kanghan Oh, and Il-Seok Oh. Regional multi-scale approach for visually pleasing explanations of deep neural networks. *arXiv e-prints*, page arXiv:1807.11720, Jul 2018. 2

[37] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3145–3153, 2017. 2

[38] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *International Conference on Learning Representations (ICLR)*, 2014. 2, 7

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv e-prints*, page arXiv:1409.1556, Sep 2014. 6

[40] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv e-prints*, page arXiv:1706.03825, Jun 2017. 2

[41] Sharon D. Solomon, Emily Chew, Elia J. Duh, Lucia Sobrin, Jennifer K. Sun, Brian L. VanderBeek, Charles C. Wykoff, and Thomas W. Gardner. Diabetic retinopathy: a position statement by the American diabetes association. *Diabetes care*, 40(3):412–418, 2017. 10, 16

[42] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 2, 7

[43] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328, 2017. 2

[44] 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 (CVPR)*, 2015. 6

[45] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 4

[46] Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, 318(22):2211–2223, 2017. 10

[47] Joanne WY Yau, Sophie L. Rogers, Ryo Kawasaki, Ecosse L. Lamoureux, Jonathan W. Kowalski, Toke Bek, Shih-Jen Chen, Jacqueline M. Dekker, Astrid Fletcher, Jakob Grauslund, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes care*, 35(3):556–564, 2012. 10

[48] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2014. 2, 7, 8

[49] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 543–559, 2016. 1, 2, 7

[50] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018. 2

[51] Yitian Zhao, Yalin Zheng, Yifan Zhao, Yonghuai Liu, Zhili Chen, Peng Liu, and Jiang Liu. Uniqueness-driven saliency analysis for automated lesion detection with applications to retinal diseases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–118. Springer, 2018. 8

[52] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object Detectors Emerge in Deep Scene CNNs. *arXiv e-prints*, page arXiv:1412.6856, Dec 2014. 2

[53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 1, 2, 8

[54] Lei Zhou, Penglin Li, Qi Yu, Yu Qiao, and Jie Yang. Automatic hemorrhage detection in color fundus images based on gradual removal of vascular branches. In *IEEE International Conference on Image Processing (ICIP)*, pages 399–403, 2016. 8