Visualization of Supervised and Self-Supervised Neural Networks via Attribution Guided Factorization

Shir Gur, Ameen Ali, Lior Wolf
The School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Abstract

Neural network visualization techniques mark image locations by their relevancy to the network's classification. Existing methods are effective in highlighting the regions that affect the resulting classification the most. However, as we show, these methods are limited in their ability to identify the support for alternative classifications, an effect we name *the saliency bias* hypothesis. In this work, we integrate two lines of research: gradient-based methods and attribution-based methods, and develop an algorithm that provides per-class explainability. The algorithm back-projects the per pixel local influence, in a manner that is guided by the local attributions, while correcting for salient features that would otherwise bias the explanation. In an extensive battery of experiments, we demonstrate the ability of our methods to class-specific visualization, and not just the predicted label. Remarkably, the method obtains state of the art results in benchmarks that are commonly applied to gradient-based methods as well as in those that are employed mostly for evaluating attribution methods. Using a new unsupervised procedure, our method is also successful in demonstrating that self-supervised methods learn semantic information. Our code is available at: https://github.com/shirgur/AGFVisualization.

Introduction

The most common class of explainability methods for image classifiers visualize the reason behind the classification of the network as a heatmap. These methods can make the rationale of the decision accessible to humans, leading to higher confidence in the ability of the classifier to focus on the relevant parts of the image, and not on spurious associations, and help debug the model. In addition to the human user, the "computer user" can also benefit from such methods, which can seed image segmentation techniques (Ahn, Cho, and Kwak 2019; Huang et al. 2018; Wang et al. 2019; Hoyer et al. 2019), or help focus generative image models, among other tasks.

The prominent methods in the field can be divided into two families: (i) gradient-based maps, which consider the gradient signal as it is computed by the conventional back-propagation approach (Sundararajan, Taly, and Yan 2017; Smilkov et al. 2017; Srinivas and Fleuret 2019; Selvaraju et al. 2017) and (ii) relevance propagation methods (Bach et al. 2015; Nam et al. 2019; Gu, Yang, and Tresp 2018; Iwana, Kuroki, and Uchida 2019), which project high-level activations back to the input domain, mostly based on the deep Taylor decomposition by (Montavon et al. 2017). The two families are used for different purposes, and are evaluated by different sets of experiments and performance metrics. As we show in Sec. , both types of methods have complementary sets of advantages and disadvantages: gradient based methods, such as Grad-CAM, are able to provide a class specific visualization for deep layers, but fail to do so for the input image, and also provide a unilateral result. In contrast, attribution based methods, excel in visualizing at the input image level, and have bipartite results, but lack in visualizing class specific explanations.

We present a novel method for class specific visualization of deep image recognition models. The method is able to overcome the limitations of the previous work, by combining ideas from both families of methods, and accumulating across the layers both gradient information, and relevance attribution. The method corrects for what we term *the saliency bias*. This bias draws the attention of the network towards the salient activations, and can prevent visualizing other image objects. This has led to the claims that visualization methods mimic the behavior of edge detectors, and that the generated heatmaps are largely independent of the network weights (Adebayo et al. 2018).

There are two different questions that explainability method often tackle: (i) which pixels affect classification the most, and (ii) which pixels are identified as belonging to the predicted class. Our method answers the second one and outperforms the relevant literature methods in multiple ways. First, the locations we identify are much more important to the classification outcome than those of the baselines, when looking inside the region of a target class. Second, we are able to identify regions of multiple classes in each image, and not just the prominent class, see Fig. 1. Our method greatly outperforms recent multi-class work (Gu, Yang, and Tresp 2018; Iwana, Kuroki, and Uchida 2019).

The main contributions of our work are: (1) a novel explainability method, that combines both gradient and attribution techniques, as well as a new attribution guided factorization technique for extracting informative class-specific attributions from the input feature map and its gradients. (2)

(a) (b) (c) (d) (e) (f)

Figure 1: Visualizations by our method for a pre-trained VGG-19. (a,d) input images. (b,e) the heatmap generated for the top label. (c,f) same for the 2nd highest prediction.

we point to and correct, for the first time, for *the saliency bias*. This bias hindered all previous attempts to explain more than the top decision, and is the underlying reason for the failures of explainability methods demonstrated in the literature. (3) state-of-the-art performance in both negative perturbation and in segmentation-based evaluation. The former is often used to evaluate attribution methods, and the latter is often used to evaluate gradient-based methods. (4) using a novel procedure, we show for the first time as far as we can ascertain, that self-supervised networks implicitly learn semantic segmentation information.

Related Work

Many explainability methods belong to one of two classes: attribution and gradient methods. Both aim to explain and visualize neural networks, but differ in their underlying concepts. A summary of these methods with their individual properties is presented in Tab. 1.

Methods outside these two classes include those that generate salient feature maps (Dabkowski and Gal 2017; Simonyan, Vedaldi, and Zisserman 2013; Mahendran and Vedaldi 2016; Zhou et al. 2016; Zeiler and Fergus 2014; Zhou et al. 2018), Activation Maximization (Erhan et al. 2009) and Excitation Backprop (Zhang et al. 2018). Extremal Perturbation methods (Fong, Patrick, and Vedaldi 2019; Fong and Vedaldi 2017) are applicable to black box models, but suffer from high computational complexity. Shapley-value based methods (Lundberg and Lee 2017), despite their theoretical appeal, are known to perform poorly in practice. Therefore, while we compare empirically with several Shaply and perturbation methods, we focus on the the newer gradient and attribution methods.

Attribution propagation methods follow the Deep Taylor Decomposition (DTD) method, of Montavon et al. (2017), which decompose the network classification decision into the contributions of its input elements. Following this line of work, methods, such as Layer-wise Relevance Propagation (LRP)by Bach et al. (2015), use DTD to propagate relevance from the predicated class, backward, to the input image, in neural networks with a rectified linear unit (ReLU) non-linearity. The PatterNet and PaternAttribution (Kindermans et al. 2017) methods yield similar results to LRP. A disadvantage of LRP, is that it is class agnostic, meaning that propagating from different classes yields the same visualization. Contrastive-LRP (CLRP) by Gu, Yang, and Tresp (2018) and Softmax-Gradient-LRP (SGLRP) by Iwana, Kuroki, and Uchida (2019) use LRP to propagate results of the target class and, in contrast to all other classes, in order to produce a class specific visualization. Nam et al. (2019) presented RAP, a DTD approach that partitions attributions to positive and negative influences, following a mean-shift approach. This approach is class agnostic, as we demonstrate in Sec. . Deep Learning Important FeaTures (DeepLIFT) (Shrikumar, Greenside, and Kundaje 2017) decomposes the output prediction, by assigning the differences of contribution scores between the activation of each neuron to its reference activation.

Gradient based methods use backpropagation to compute the gradients with respect to the layer's input feature map, using the chain rule. The Gradient*Input method by Shrikumar et al. (2016) computes the (signed) partial derivatives of the output with respect to the input, multiplying it by the input itself. Integrated Gradients (Sundararajan, Taly, and Yan 2017), similar to Shrikumar et al. (2016), computes the multiplication of the inputs with its derivatives, only they compute the average gradient while performing a linear interpolation of the input, according to some baseline that is defined by the user. SmoothGrad by Smilkov et al. (2017), visualize the mean gradients of the input, while adding to the input image a random Gaussian noise at each iteration. The FullGrad method by Srinivas and Fleuret (2019) suggests computing the gradients of each layer's input, or bias, followed by a post-processing operator, usually consisting of the absolute value, reshaping to the size of the input, and summing over the entire network. Where for the input layer, they also multiply by the input image before post-processing. As we show in Sec. , FullGrad produces class agnostic visualizations. On the other hand, Grad-CAM by Selvaraju et al. (2017) is a class specific approach, combining both the input features, and the gradients of a network's layer. This approach is commonly used in many applications due to this property, but its disadvantage rests in the fact that it can only produce results for very deep layers, resulting in coarse visualization due to the low spatial dimension of such deep layers.

Propagation Methods

We define the building blocks of attribution propagation and gradient propagation that are used in our method.

Attribution Propagation: Let $x^{(n)}$, $\theta^{(n)}$ be the input feature map and weights of layer $L^{(n)}$, respectively, where $n \in [1 \dots N]$ is the layer index in a network, consisting of

	Int. Grad	Smooth Grad	LRP	LRP$_{\alpha\beta}$	Full Grad	Grad CAM	RAP	CLRP	SGLRP	Ours
Gradients	✓	✓			✓	✓				✓
Attribution			✓	✓			✓	✓	✓	✓
Class Specific						✓		✓	✓	✓
Input Domain	✓	✓	✓	✓	✓		✓	✓	✓	✓

Table 1: Properties of various visualization methods, which can be largely divided into gradient based and attribution based. Most methods are not class specific, and except for Grad-CAM, all methods project all the way back to the input image.

N layers. In this field's terminology, layer $n-1$ is downstream of layer n, layer N processes the input, while layer 1 produces the final output.

Let $x^{(n-1)} = L^{(n)}(x^{(n)}, \theta^{(n)})$ to be the result of applying layer $L^{(n)}$ on $x^{(n)}$. The relevancy of layer $L^{(n)}$ is given by $R^{(n)}$ and is also known as the attribution.

Definition 1 (Montavon et al. 2017) *The* generic attribution propagation rule *is defined, for two tensors,* $\mathbf{X}$ *and* Θ*, as:*

$$R_j^{(n)} = \mathcal{G}_j^{(n)}(\mathbf{X}, \Theta, R^{(n-1)}) \tag{1}$$
$$= \sum_i \mathbf{X}_j \frac{\partial L_i^{(n)}(\mathbf{X}, \Theta)}{\partial \mathbf{X}_j} \frac{R_i^{(n-1)}}{L_i^{(n)}(\mathbf{X}, \Theta)}$$

Typically, $\mathbf{X}$ is related to the layer's input $x^{(n)}$, and Θ to its weights $\theta^{(n)}$. LRP (Binder et al. 2016) can be written in this notation by setting $\mathbf{X} = x^{(n)+}$ and $\Theta = \theta^{(n)+}$, where $\tau^+ = \max(0, \tau)$ for a tensor τ. Note that Def. 1 satisfies the *conservation rule*:

$$\sum_j R_j^{(n)} = \sum_i R_i^{(n-1)} \tag{2}$$

Let y be the output vector of a classification network with $\mathcal{C}$ classes, and let y^t represent the specific value of class $t \in \mathcal{C}$. LRP defines $R^{(0)} \in \mathbb{R}^{|\mathcal{C}|}$ to be the a zeros vector, except for index t, where $R_t^{(0)} = y^t$. Similarly, the CLRP (Gu, Yang, and Tresp 2018) and SGLRP (Iwana, Kuroki, and Uchida 2019) methods calculate the difference between two LRP results, initialized with two opposing $R^{(0)}$ for "target" and "rest", propagating relevance from the "target" class, and the "rest" of the classes, *e.g.*CLRP is defined as:

$$CLRP = R_{tgt}^{(N)} - \mathcal{N}\big(R_{rst}^{(N)}, R_{tgt}^{(N)}\big) \tag{3}$$

where $R_{tgt}^{(0)} = R^{(0)}$, $R_{rst}^{(0)} = (y - R^{(0)})/(|\mathcal{C}| - 1)$, and $\mathcal{N}$ is a normalization term $\mathcal{N}(a, b) = a\frac{\sum b}{\sum a}$.

Δ-Shift: Def. 1 presented a generic propagation rule that satisfies the conservation rule in Eq. 2. However, in many cases, we would like to add a residual signal denoting another type of attribution. The Δ-Shift corrects for the deviation from the conservation rule.

Definition 2 *Given a generic propagation result* $\mathcal{G}^{(n)}$*, following Eq. 2, and a residual tensor* $\mathbf{r}^{(n)}$*, the* Δ*-Shift is defined as follows:*

$$\Delta_{\text{shift}}^{(n)}(\mathcal{G}^{(n)}, \mathbf{r}^{(n)}) = \mathcal{G}^{(n)} + \mathbf{r}^{(n)} - \frac{\sum \mathbf{r}^{(n)}}{\sum \mathbb{1}_{\mathcal{G}^{(n)} \neq 0}} \tag{4}$$

Note that we divide the sum of the residual signal by the number of non-zero neurons. While not formulated this way, the RAP method (Nam et al. 2019) employs this type of correction defined in Def. 2.

Gradient Propagation: The propagation in a neural network is defined by the chain rule.

Definition 3 *Let* $\mathcal{L}$ *be the loss of a neural network. The input feature gradients,* $x^{(n)}$ *of layer* $L^{(n)}$*, with respect to* $\mathcal{L}$ *are defined by the chain rule as follows:*

$$\nabla x_j^{(n)} := \frac{\partial \mathcal{L}}{\partial x_j^{(n)}} = \sum_i \frac{\partial \mathcal{L}}{\partial x_i^{(n-1)}} \frac{\partial x_i^{(n-1)}}{\partial x_j^{(n)}} \tag{5}$$

Methods such as FullGrad (Srinivas and Fleuret 2019) and SmoothGrad (Smilkov et al. 2017) use the raw gradients, as defined in Eq.5, for visualization. Grad-CAM (Selvaraju et al. 2017), on the other hand, performs a weighted combination of the input feature gradients, in order to obtain a class specific visualization, defined as follows:

$$\text{Grad-CAM}\big(x^{(n)}, \nabla x^{(n)}\big) = \left(\frac{1}{|C|} \sum_{c \in [C]} x_c^{(n)} \sum_{\substack{h \in [H] \\ w \in [W]}} \nabla x_{c,h,w}^{(n)}\right)^+ \tag{6}$$

where $\nabla x_{c,h,w}^{(n)}$ is the specific value of the gradient C-channel tensor $x^{(n)}$ at channel c and pixel (h, w), and $x_c^{(n)}$ is the entire channel, which is a matrix of size $H \times W$.

Guided Factorization: The explanation should create a clear separation between the positively contributing regions, or the foreground, and the negatively contributing ones, referred to as the background. This is true for both the activations and the gradients during propagation. Ideally, the relevant data would be partitioned into two clusters — one for positive contributions and one for the negative contributions. We follow the partition problem (Yuan, Wang, and Cheriyadat 2015; Gao et al. 2016), in which the data is divided spatially between positive and negative locations, in accordance with the sign of a partition map $\phi \in \mathbb{R}^{H \times W}$.

Given a tensor $\mathbf{Y} \in \mathbb{R}^{C \times H \times W}$, we re-write it as a matrix in the form of $\mathbf{Y} \in \mathbb{R}^{C \times HW}$. We compute the Heaviside function of $\mathbf{Y}$ using a sigmoid function: $\mathbf{H} = \text{sigmoid}(\mathbf{Y})$. The matrix $H \in [0, 1]^{C \times HW}$ is a positive-matrix, and we consider the following two-class non-negative matrix factorization $\mathbf{H} = \mathbf{RW}$, where $\mathbf{W} \in \mathbb{R}^{+\,2 \times HW}$ contains the spatial mixing weights, and the representative matrix

$\mathbf{R} = [\mathbf{R}_b\, \mathbf{R}_f]$, defined by the mean of each class in the data tensor $\mathbf{H}$ based on the assignment of ϕ:

$$\mathbf{R}_c^f = \frac{\sum_i^{HW} \mathbf{H}(\mathbf{Y})_{c,i} \odot \mathbb{1}_{\phi_i > 0}}{\sum_i^{HW} \mathbb{1}_{\phi_i > 0}}, \mathbf{R}_c^b = \frac{\sum_i^{HW} \mathbf{H}(\mathbf{Y})_{c,i} \odot \mathbb{1}_{\phi_i \leqslant 0}}{\sum_i^{HW} \mathbb{1}_{\phi_i \leqslant 0}}$$

where $\mathbf{R}^f, \mathbf{R}^b \in \mathbb{R}^{+C}$, $c \in C$ is the channel dimension and $\odot$ denotes the Hadamard product.

We estimate the matrix $\mathbf{W}$ of positive weights by least squares $\mathbf{W} = [\mathbf{W}_b\, \mathbf{W}_f] = ((\mathbf{R}^\top \mathbf{R})^{-1} \mathbf{R}^\top \mathbf{H})^+$, where $\mathbf{W}_f, \mathbf{W}_b \in \mathbb{R}^{+HW}$. Combining the foreground weights W_f with the background weights W_b into the same axis is done by using both negative and positive values, leading to the following operator: $\mathcal{F}(\mathbf{Y}, \phi) = \mathbf{W}_f - \mathbf{W}_b$, where $\mathcal{F}$ is a function that takes $\mathbf{Y}$ and ϕ, as inputs. We further normalize $\mathbf{Y}$ using $\mathcal{N}_{\max}(a) = \frac{a}{\max(a)}$, to allow multiple streams to be integrated together:

$$\bar{\mathcal{F}}(\mathbf{Y}, \phi) = \mathcal{F}(\mathcal{N}_{\max}(\mathbf{Y}), \phi) \tag{7}$$

The Integrated Method

Let M be a multiclass CNN classifier ($\mathcal{C}$ labels), and $I = x^{(N)}$ be the input image. The network M outputs a score vector $y \in \mathbb{R}^{|\mathcal{C}|}$, obtained before applying the $\mathrm{softmax}$ operator. Given any target class t, our goal is to explain where (spatially) in I lies the support for class t. The method is composed of two streams, gradients and attribution propagation. Each step, we use the previous values of the two, and compute the current layer's input gradient and attribution.

There are three major components to our method, (i) propagating attribution results using Def. 1, (ii) factorizing the activations and the gradients in a manner that is guided by the attribution, and (iii) performing attribution aggregation and shifting the values, such that the conservation rule is preserved. The shift splits the neurons into those with a positive and negative attribution. The complete algorithm is listed as part of the supplementary material.

Initial Attribution Propagation

As shown in (Gu, Yang, and Tresp 2018; Iwana, Kuroki, and Uchida 2019), the use of different initial relevance can result in improved results. Let $\Phi^{(n)}$ and $\Phi^{(n-1)}$ be the output and input class attribution maps of layer $L^{(n)}$, respectively. We employ the following initial attribution for explaining decision t. Let $y := x^{(0)}$ be the output vector of the classification network (logits), we compute the initial attribution $\Phi^{(1)}$ as:

$$\hat{y} = \mathrm{softmax}\left(y^t \exp\left(-\frac{1}{2}\left(\frac{y - y^t}{\max \|y - y^t\|_1}\right)^2\right)\right)$$
$$\frac{\partial \hat{y}^t}{\partial x_j^{(1)}} = \sum_i \frac{\partial \hat{y}^t}{\partial y_i} \frac{\partial y_i}{\partial x_j^{(1)}}, \Phi^{(1)} = x^{(1)} \odot \frac{\partial \hat{y}^t}{\partial x^{(1)}} \tag{8}$$

In this formulation, we replace the psuedo-probabilities of vector y with another vector, in which class t that we wish to provide an explanation for is highlighted, and the rest of the classes are scored by the closeness of their assigned probability to that of t. This way, the explanation is no longer dominated by the predicted class.

Class Attribution Propagation

As mentioned above, there are two streams in our method. The first propagates the gradients, which are used for the factorization process we discuss in the next section, and the other is responsible for propagating the resulting class attribution maps $\Phi^{(n-1)}$, using DTD-type propagations. This second stream is more involved, and the computation of $\Phi^{(n)}$ has multiple steps.

The first step is to propagate $\Phi^{(n-1)}$ through $L^{(n)}$ following Eq. 1, using two variants. Both variants employ $\Phi^{(n-1)}$ as the tensor to be propagated, but depending on the setting of $\mathbf{X}$ and Θ, result in different outcomes. The first variant considers the absolute influence $\mathbf{C}^{(n)}$, defined by:

$$\mathbf{C}^{(n)} = \mathcal{G}^{(n)}(|x^{(n)}|, |\theta^{(n)}|, \Phi^{(n-1)}) \tag{9}$$

The second variant computes the input-agnostic influence $\mathbf{A}^{(n)}$, following Eq. 1:

$$\mathbf{A}^{(n)} = \mathcal{G}^{(n)}(\mathbb{1}, |\theta^{(n)}|, \Phi^{(n-1)}) \tag{10}$$

where $\mathbb{1}$ is an all-ones tensor of the shape of $x^{(n)}$. We choose the input-agnostic propagation because features in shallow layers, such as edges, are more local and less semantic. It, therefore, reduces the sensitivity to texture.

Residual Update

As part of the method, we compute in addition to $\mathbf{C}^{(n)}$, the factorization of both the input feature map of layer $L^{(n)}$ and its gradients. This branch is defined by the chain rule in Eq. 5, where we now consider $\mathcal{L} = \hat{y}^t$. The factorization results in *foreground* and *background* partitions, using guidance from $\mathbf{C}^{(n)}$. This partition follows the idea of our attribution properties, where positive values are part of class t, and negatives otherwise. We, therefore, employ the following attribution guided factorization (Eq. 7 with respect to $x^{(n)}$ and $\nabla x^{(n)}$) as follows:

$$\mathbf{F}_x^{(n)} = \bar{\mathcal{F}}\big(x^{(n)}, \mathbf{C}^{(n)}\big)^+, \quad \mathbf{F}_{\nabla x}^{(n)} = \bar{\mathcal{F}}\big(\nabla x^{(n)}, \mathbf{C}^{(n)}\big)^+ \tag{11}$$

note that we only consider the positive values of the factorization update, and that the two results are normalized by their maximal value. Similarly to (Sundararajan, Taly, and Yan 2017; Selvaraju et al. 2017), we define the input-gradient interaction:

$$\mathbf{M}_{x\nabla x}^{(n)} = \mathcal{N}_{\max}\left(\left(\frac{1}{|C|}\sum_{c \in [C]}\left(x_c^{(n)} \odot \nabla x_c^{(n)}\right)\right)^+\right) \tag{12}$$

The residual attribution is then defined by all attributions other than $\mathbf{C}^{(n)}$:

$$\mathbf{r}^{(n)} = \mathbf{A}^{(n)} + \mathbf{F}_{\nabla x}^{(n)} + \frac{\mathbf{F}_x^{(n)} + \mathbf{M}_{x\nabla x}^{(n)}}{1 + \exp(-\mathbf{C}^{(n)})} \tag{13}$$

We observe that both $\mathbf{F}_x^{(n)}$ and $\mathbf{M}_{x\nabla x}^{(n)}$ are affected by the input feature map, resulting in the *saliency bias* effect (see Sec.). As a result, we penalize their sum according to $\mathbf{C}^{(n)}$, in a manner that emphasises positive attribution regions.

Algorithm 1 Class Attribution Propagation with Guided Factorization.

Require: M : Neural network model, $I \in \mathbb{R}^{C\times H\times W}$: Input images, $t \in \mathcal{C}$: Target class.

$M(I)$ ▷ Forward-pass, save intermediate feature maps
$\sigma \leftarrow \max \|y - y^t\|_1$
$\hat{y} \leftarrow \text{softmax}\left(y^t \exp\left(-\frac{1}{2}\left(\frac{y-y^t}{\sigma}\right)^2\right)\right)$
$\Phi^{(1)} \leftarrow x^{(1)} \odot \frac{\partial \hat{y}^t}{\partial x^{(1)}}$ ▷ Initial Attribution - First Linear layer
for linear layers, $n > 1$ **do**
  $\mathbf{C}_j^{(n)} \leftarrow \mathcal{G}^{(n)}(|x^{(n)}|, |\theta^{(n)}|, \Phi^{(n-1)})$ ▷ Absolute influence
  if next layer is 2D **then**
    Reshape x and ∇x to the previous 2D form
    $\mathbf{M}^{(n)} \leftarrow \mathcal{N}_{\max}\left(\left(\frac{1}{|C|}\sum_{c\in[C]}\left(x_c^{(n)} \odot \nabla x_c^{(n)}\right)\right)^+\right)$
    $\mathbf{r}^{(n)} \leftarrow \mathbf{M}^{(n)}$ ▷ Residual
  else
    $\mathbf{r}^{(n)} \leftarrow 0$
  $\Phi^{(n)} \leftarrow \Delta_{shift}^{(n)}(\mathbf{C}^{(n)}, \mathbf{r}^{(n)})$ ▷ Shifting by the residual
for convolution layers **do**
  Compute $\mathbf{C}^{(n)}$, $\mathbf{M}^{(n)}$
  $\mathbf{A}_j^{(n)} \leftarrow \mathcal{G}^{(n)}(\mathbb{1}, |\theta^{(n)}|, \Phi^{(n-1)})$ ▷ Input agnostic attribution
  $\mathbf{F}_x^{(n)} \leftarrow \bar{\mathcal{F}}(x^{(n)}, \mathbf{C}^{(n)})^+$ ▷ Factorization of input feature map
  $\mathbf{F}_{\nabla x}^{(n)} \leftarrow \bar{\mathcal{F}}(\nabla x^{(n)}, \mathbf{C}^{(n)})^+$ ▷ Factorization of input feature map gradients
  $\mathbf{r}^{(n)} \leftarrow \mathbf{A}^{(n)} + \mathbf{F}_{\nabla x}^{(n)} + \frac{\mathbf{F}_x^{(n)} + \mathbf{M}_{x\nabla x}^{(n)}}{1+\exp(-\mathbf{C}^{(n)})}$ ▷ Residual
  Compute $\Phi^{(n)}$

Algorithm 2 self-supervised explainability by adopting nearest neighbors

Require: I input image, $\mathcal{S}$ set of images, $M = \phi_\tau \circ \phi_F$ the SSL network, where ϕ_F extracts the features and ϕ_τ is the linear SSL classifier.

$L_I = \phi_F(I)$ ▷ L_I is the latent vector of image I
$\mathcal{L} \leftarrow \{\phi_F(J) | J \in \mathcal{S}\}$ ▷ $\mathcal{L}$ is the set of all latent vectors for all images in $\mathcal{S}$
$L_N = \arg\min_{L\in\mathcal{L}} \|L_I - L\|$ ▷ L_N is the nearest neighbor of L
$S_I = L_I - L_N$ ▷ Subtracting L_N from L_I to emphasize the unique elements of L_I
$v = \phi_\tau(S_I)$ ▷ Forward pass with the new latent vector
$t = \arg\max v$ ▷ Choose the class with the highest probability
Apply Alg. 1, with the input tuple M, I, t

We note that $\sum_j(\mathbf{C}_j^{(n)} + \mathbf{r}_j^{(n)}) \neq \sum_i \Phi_i^{(n-1)}$, and the residual needs to be compensated for, in order to preserve the conservation rule. Therefore, we perform a Δ-shift as defined in Def. 2, resulting in the final attribution:

$$\Phi^{(n)} = \Delta_{shift}^{(n)}(\mathbf{C}^{(n)}, \mathbf{r}^{(n)}) \tag{14}$$

The full algorithm of our method is presented in Alg. 1.

Explaining Self-Supervised Learning (SSL)

SSL is proving to be increasingly powerful and greatly reduces the need for labeled samples. However, no explainability method was applied to verify that these models, which are often based on image augmentations, do not ignore localized image features.

Since no label information is used, we rely on the classifier of the self-supervised task itself, which has nothing to do with the classes of the datasets. We consider for each image, the image that is closest to it in the penultimate layer's activations. We then subtract the logits of the self supervised task of the image to be visualized and its nearest neighbor, to emphasize what is unique to the current image, and then use explainability methods on the predicted class of the self-supervised task. The full algorithm for SSL visualization is presented in Alg. 2.

Experiments

Qualitative Evaluation: Fig. 2(a,b) present sample visualization on a representative set of images for networks trained on ImageNet, using VGG19 and ResNet-50, respectively. In these figures, we visualize the top-predicted class. More results can be found in the supplementary, including the output for the rest of the methods, which are similar to other methods and are removed for brevity.

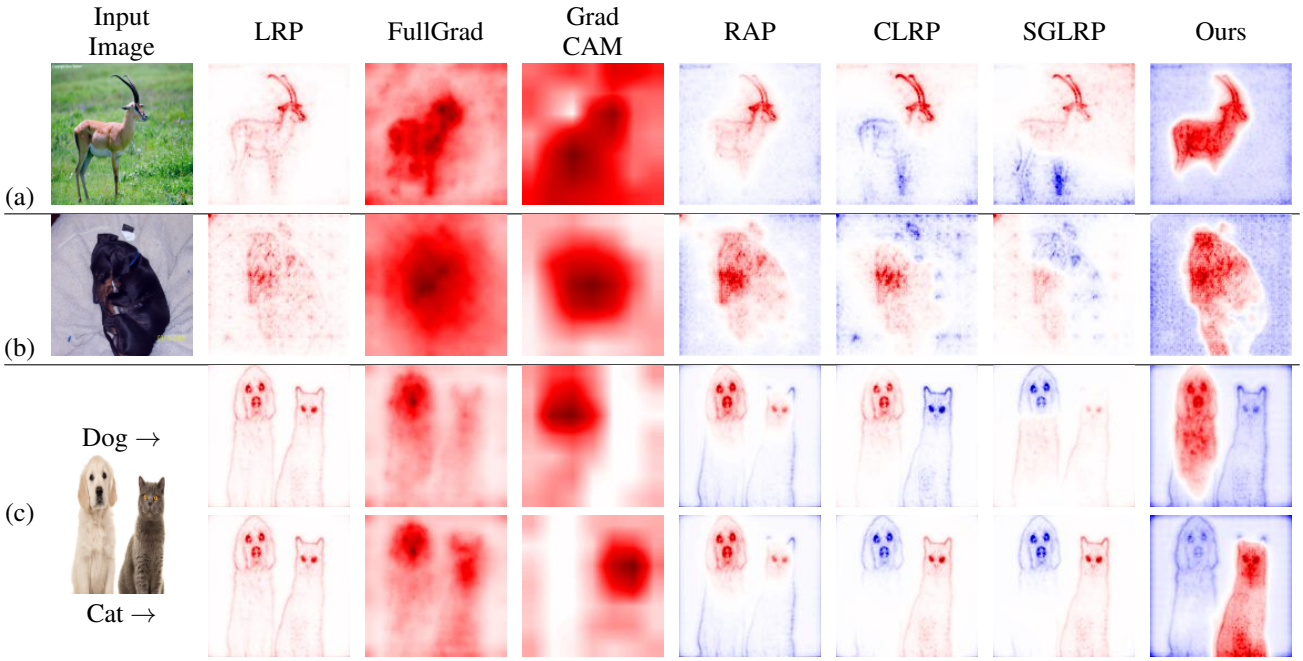


Figure 2: (a) A comparison between methods for VGG-19. (b) Same for ResNet-50. (c) Visualization of two different classes for VGG-19. Many more results can be found in the supplementary material.

	Int. Grad	Smooth Grad	LRP	LRP$_{\alpha\beta}$	Full Grad	Grad CAM	RAP	CLRP	SGLRP	Ours
Predicted	10.4	12.0	26.8	18.8	32.1	37.8	38.7	29.8	29.9	**38.9**
Target	10.5	12.1	26.8	18.8	32.4	39.4	38.7	31.6	32.4	**40.0**

Table 2: Area Under the Curve (AUC) for the two negative perturbation tests, showing results for predicted and target class. The class-agnostic methods either perform worse or experience insignificant change on the target class test.

The preferable visualization quality provided by our method is strikingly evident. One can observe that (i) LRP, FullGrad and Grad-CAM output only positive results, wherein LRP edges are most significant, and in all three, the threshold between the object and background is ambiguous. (ii) CLRP and SGLRP, which apply LRP twice, have volatile outputs. (iii) RAP is the most consistent, other than ours, but falls behind in object coverage and boundaries. (iv) Our method produces relatively complete regions with clear boundaries between positive and negative regions.

In order to test whether each method is class-agnostic or not, we feed the classifier images containing two clearly seen objects, and propagate each object class separately. In Fig. 2(c) we present results for a sample image. As can be seen, LRP, FullGrad and RAP output similar visualizations for both classes. Grad-CAM, on the other hand, clearly shows a coarse region of the target class, but lacks the spatial resolution. CLRP and SGLRP both achieve class separation, and yet, they are highly biased toward image edges, and do not present a clear separation between the object and its background. Our method provides the clearest visualization, which is both highly correlated with the target class, and is less sensitive toward edges. More samples can be found in the supplementary.

Quantitative Experiments: We employ two experiment settings that are used in the literature, negative perturbation and segmentation tests. We evaluate our method using three common datasets: (i) the validation set of ImageNet (Russakovsky et al. 2015) (ILSVRC) 2012, consisting of 50K images from 1000 classes, (ii) an annotated subset of ImageNet called ImageNet-Segmentation (Guillaumin, Küttel, and Ferrari 2014), containing 4,276 images from 445 categories, and (iii) the PASCAL-VOC 2012 dataset, depicting 20 foreground object classes and one background class, and containing 10,582 images for training, 1449 images for validation and 1,456 images for testing.

Negative Perturbation Experiments: The negative perturbation test is composed of two stages, first, a pre-trained network is used to generate the visualizations of the ImageNet validation set. In our experiments, we use the VGG-19 architecture, trained on the full ImageNet training set. Second, we mask out an increasing portion of the image, starting from lowest to highest values, determined by the explainability method. At each step, we compute the mean accuracy of the

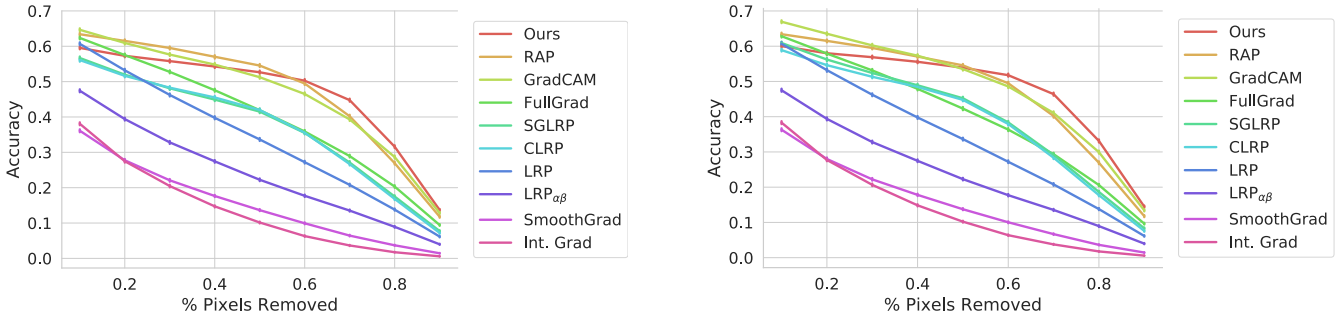


Figure 3: Negative perturbation results on the ImageNet validation set of **(a)** the predicted and **(b)** the target class. Show is the change in accuracy, when removing a fraction of the image according to the attribution value, starting from lowest to highest.

Dataset	Metric	Gradient SHAP	Meaningful Perturbation	Int. Grad	Smooth Grad	LRP	LRP$_{\alpha\beta}$	Full Grad	Grad CAM	RAP	CLRP	SGLRP	Ours
ImageNet	Pixel Acc.	49.9	29.6	72.7	70.1	74.9	29.6	75.8	66.5	78.8	53.3	54.5	**79.9**
	mAP	50.0	45.4	67.4	65.1	69.9	47.5	70.6	62.4	73.7	54.6	55.1	**76.7**
VOC'12	Pixel Acc.	9.3	10.2	70.1	69.9	66.6	70.8	26.9	52.6	72.8	72.8	73.1	**77.5**
	mAP	35.2	22.7	34.9	34.3	40.8	37.0	19.7	41.7	39.6	37.9	32.6	**45.9**

Table 3: Quantitative segmentation results on (a) ImageNet and (b) PASCAL-VOC 2012.

pre-trained network. We repeat this test twice: once for the explanation of the top-1 predicted class, and once for the ground truth class.

The results are presented in Fig. 3 and Tab. 2. As can be seen, our method achieves the best performance across both tests, where the margin is highest when removing $40\% - 80\%$ of the pixels.

Semantic Segmentation Metrics: To evaluate the segmentation quality obtained by each explainability method, we compare each to the ground truth segmentation maps of the ImageNet-Segmentation dataset, and the PASCAL-VOC 2012, evaluating by pixel-accuracy and mean average-precision. We follow the literature benchmarks for explainability that are based on labeled segments (Nam et al. 2019). Note that these are not meant to provide a full weakly-supervised segmentation solution, which is often obtained in an iterative manner. Rather, its goal is to demonstrate the ability of each method without follow-up training.

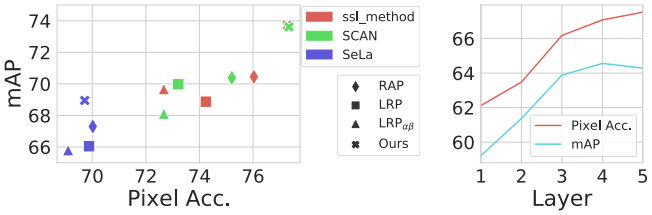
For the first dataset, we employed the pre-trained VGG19 classifier trained on ImageNet training set, and compute the explanation for the top-predicted class (we have no access to the ground truth class) and compare it to the ground truth mask provided in the dataset. For the second, we trained a multi-label classifier on the PASCAL-VOC 2012 training set, and consider labels with a probability larger than 0.5 to extract the explainability maps.

For methods that provide both positive and negative values (Integrated Grad, Gradient SHAP, LRP$_{\alpha\beta}$, RAP, CLRP, SGLRP, and ours), we consider the positive part as the segmentation map of that object. For methods that provide only positive values (Smooth Grad, Full Grad, GradCAM, LRP, Meaningful Perturbation), we threshold the obtained maps at the mean value to obtain the segmentation map. Results are reported in Tab. 3, demonstrating a clear advantage of our method over all nine baseline methods, for all datasets and metrics. Other methods seem to work well only in one of the datasets or present a trade-off between the two metrics.

Explainability for Self-Supervised Models: We use three state-of-the-art SSL models: ResNet-50 trained with either SCAN (Van Gansbeke et al. 2020) or SeLa (Asano, Rupprecht, and Vedaldi 2019b), and an Alexnet by Asano, Rupprecht, and Vedaldi (2019a), which we denote as RotNet.

Fig. 4(a) shows the segmentation performance for the ImageNet ground truth class of the completely unsupervised SSL methods (color) using different explainability methods (shape). For all models our explainability method outperforms the baselines in both mAP and pixel accuracy, except for RotNet where the mAP is considerably better and the pixel accuracy is slightly lower. Fig. 4(b) shows the increase of segmentation performance for RotNet with our method, as we visualize deeper layers of SSL Alexnet, using a supervised linear post-training as proposed by Asano, Rupprecht, and Vedaldi (2019a). This finding is aligned with the classification results by a single linear layer of Asano, Rupprecht, and Vedaldi (2019a), which improve with the layer index.

Tab. 4 compares RAP (which seems to be the strongest baseline) and our method in the predicted SSL class (not ImageNet) negative perturbation setting. Evidently, our method has superior performance. SSL results also seem correlated with the fully supervised ones (note that the architecture and the processing of the fully connected layer is different from the one used in Tab. 2). Two alternatives to our novel SSL



(a) (b)

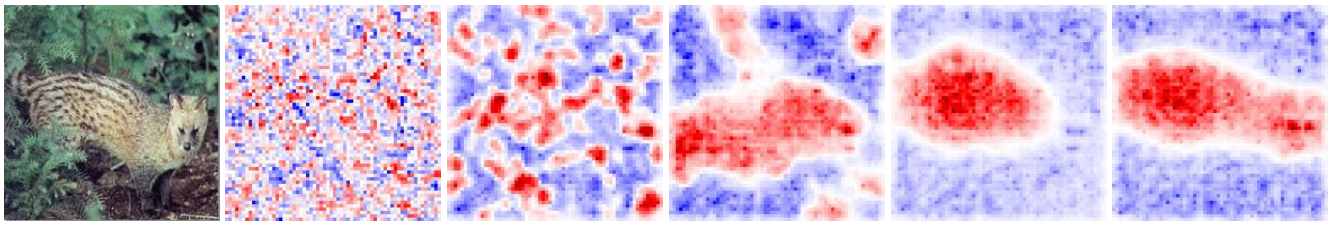
Figure 4: Quantitative results for self-supervised methods in the segmentation task. **(a)** Comparison of different explainability methods for SCAN, SeLa and RotNet. **(b)** Per-layer performance of RotNet using linear-probes.

	Ours			RAP		
Method	(Δ)	($\sum$)	w/o	(Δ)	($\sum$)	w/o
Supervised	**42.4**	42.1	41.2	41.6	41.2	40.0
SeLa	**37.8**	37.4	36.9	37.2	36.3	34.5
SCAN	**38.1**	37.9	37.0	37.4	36.8	34.8

Table 4: AUC for negative perturbation tests for self-supervised methods - SeLa and SCAN. (Δ) is our method; ($\sum$) is an alternative that adds instead of subtracts, w/o does not consider the neighbor at all. RAP, which is the best baseline in Tab. 2, is used as a baseline.

	Ours	Only $\mathbf{C}^{(n)}$	without					$\mathbf{r}^{(n)}$ = Grad-CAM
			$\mathbf{A}^{(n)}$	$\mathbf{F}_x^{(n)}$	$\mathbf{F}_{\nabla x}^{(n)}$	$\mathbf{M}^{(n)}$	$(1+\exp(-\mathbf{C}^{(n)}))^{-1}$	
AUC	**38.9**	38.2	34.6	37.5	37.1	37.2	37.5	37.1

Table 5: AUC results in the negative perturbation test for variations of our method.



(a) (b) (c) (d) (e) (f)

Figure 5: The visualization after each layer. (a) original image. (b) visualization after first layer. (c) visualization after second layer. (d) visualization after third layer. (e) visualization after the fourth layer. (e) visualization after last layer.

procedure are also presented: in one, the difference from the neighbor is replaced with a sum. In the other, no comparison to the neighbor takes place. Our procedure for SSL explainability is superior for both our method and RAP.

Fig. 5 shows the visualization obtained for the input image, when training a supervised linear classifier after each layer, and applying our method. As can be seen, deeper layers learn more semantic representations. Additional results are included in the supplementary.

Ablation Study: By repeatedly employing normalization, our method is kept parameter-free. In Tab. 5, we present negative perturbation results for methods that are obtained by removing one components out of our complete method. We also present the results of a similar method in which the guided attribution based residual-term $\mathbf{r}$ is replaced by a GradCam term. As can be seen, each of these modifications damages the performance to some degree. Without any residual term, the method is slightly worse than RAP, while a partial residual term further hurts performance. In the supplementary material, we show visual results of the different components and variations of our method and present observations on the contribution of each part to the final outcome.

Conclusions

Explainability plays a major rule in debugging neural networks, creating trust in the predictive capabilities of networks beyond the specific test dataset, and in seeding downstream methods that analyze the images spatially. Previous visualization methods are either class-agnostic, low-resolution, or neglect much of the object's region, focusing on edges. The separation between relevant and irrelevant image parts provided by the previous methods is also often blurry. In this work, we present a novel explainability method that outputs class-dependent explanations that are clearer and more exact than those presented by the many existing methods tested. The new method is based on combining concepts from the two major branches of the current literature: attribution methods and gradient methods. This combination is done, on equal grounds, through the usage of a non-negative matrix factorization technique that partitions the image into foreground and background regions. Finally, we propose a novel procedure for evaluating the explainability of SSL methods.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant ERC CoG 725974). The contribution of the first author is part of a Ph.D. thesis research conducted at Tel Aviv University.

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 9505–9515.

Ahn, J.; Cho, S.; and Kwak, S. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2209–2218.

Asano, Y. M.; Rupprecht, C.; and Vedaldi, A. 2019a. A critical analysis of self-supervision, or what we can learn from a single image. *arXiv preprint arXiv:1904.13132* .

Asano, Y. M.; Rupprecht, C.; and Vedaldi, A. 2019b. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371* .

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7): e0130140.

Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K.-R.; and Samek, W. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, 63–71. Springer.

Dabkowski, P.; and Gal, Y. 2017. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, 6970–6979.

Erhan, D.; Bengio, Y.; Courville, A.; and Vincent, P. 2009. Visualizing higher-layer features of a deep network. *University of Montreal* 1341(3): 1.

Fong, R.; Patrick, M.; and Vedaldi, A. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2950–2958.

Fong, R. C.; and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, 3429–3437.

Gao, M.; Chen, H.; Zheng, S.; and Fang, B. 2016. A factorization based active contour model for texture segmentation. In *2016 IEEE International Conference on Image Processing (ICIP)*, 4309–4313. IEEE.

Gu, J.; Yang, Y.; and Tresp, V. 2018. Understanding individual decisions of cnns via contrastive backpropagation. In *Asian Conference on Computer Vision*, 119–134. Springer.

Guillaumin, M.; Küttel, D.; and Ferrari, V. 2014. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision* 110(3): 328–348.

Hoyer, L.; Munoz, M.; Katiyar, P.; Khoreva, A.; and Fischer, V. 2019. Grid saliency for context explanations of semantic segmentation. In *Advances in Neural Information Processing Systems*, 6462–6473.

Huang, Z.; Wang, X.; Wang, J.; Liu, W.; and Wang, J. 2018. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7014–7023.

Iwana, B. K.; Kuroki, R.; and Uchida, S. 2019. Explaining Convolutional Neural Networks using Softmax Gradient Layer-wise Relevance Propagation. *arXiv preprint arXiv:1908.04351* .

Kindermans, P.-J.; Schütt, K. T.; Alber, M.; Müller, K.-R.; Erhan, D.; Kim, B.; and Dähne, S. 2017. Learning how to explain neural networks: Patternnet and patternattribution. *arXiv preprint arXiv:1705.05598* .

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 4765–4774.

Mahendran, A.; and Vedaldi, A. 2016. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision* 120(3): 233–255.

Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; and Müller, K.-R. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition* 65: 211–222.

Nam, W.-J.; Gur, S.; Choi, J.; Wolf, L.; and Lee, S.-W. 2019. Relative Attributing Propagation: Interpreting the Comparative Contributions of Individual Units in Deep Neural Networks. *arXiv preprint arXiv:1904.00605* .

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3): 211–252.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3145–3153. JMLR. org.

Shrikumar, A.; Greenside, P.; Shcherbina, A.; and Kundaje, A. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713* .

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

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* .

Srinivas, S.; and Fleuret, F. 2019. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems*, 4126–4135.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3319–3328. JMLR. org.

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. SCAN: Learning to Classify Images without Labels. In *European Conference on Computer Vision (ECCV)*.

Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; and Chen, X. 2019. Self-supervised Scale Equivariant Network for Weakly Supervised Semantic Segmentation. *arXiv preprint arXiv:1909.03714* .

Yuan, J.; Wang, D.; and Cheriyadat, A. M. 2015. Factorization-based texture segmentation. *IEEE Transactions on Image Processing* 24(11): 3488–3497.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

Zhang, J.; Bargal, S. A.; Lin, Z.; Brandt, J.; Shen, X.; and Sclaroff, S. 2018. Top-down neural attention by excitation backprop. *International Journal of Computer Vision* 126(10): 1084–1102.

Zhou, B.; Bau, D.; Oliva, A.; and Torralba, A. 2018. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence* .

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.