This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Occlusion Sensitivity Analysis with Augmentation Subspace Perturbation in Deep Feature Space

Pedro H. V. Valois
University of Tsukuba
pedro@cvlab.cs.tsukuba.ac.jp

Koichiro Niinuma
Fujitsu Research of America
kniinuma@fujitsu.com

Kazuhiro Fukui
University of Tsukuba
kfukui@cs.tsukuba.ac.jp

Abstract

Deep Learning of neural networks has gained prominence in multiple life-critical applications like medical diagnoses and autonomous vehicle accident investigations. However, concerns about model transparency and biases persist. Explainable methods are viewed as the solution to address these challenges. In this study, we introduce the Occlusion Sensitivity Analysis with Deep Feature Augmentation Subspace (OSA-DAS), a novel perturbation-based interpretability approach for computer vision. While traditional perturbation methods make only use of occlusions to explain the model predictions, OSA-DAS extends standard occlusion sensitivity analysis by enabling the integration with diverse image augmentations. Distinctly, our method utilizes the output vector of a DNN to build low-dimensional subspaces within the deep feature vector space, offering a more precise explanation of the model prediction. The structural similarity between these subspaces encompasses the influence of diverse augmentations and occlusions. We test extensively on the ImageNet-1k, and our class- and model-agnostic approach outperforms commonly used interpreters, setting it apart in the realm of explainable AI.

1. Introduction

Interpretability in deep learning provides insights into the complex operations of deep neural networks (DNNs), which often seem like "black boxes" due to their intricate structures. There's a growing demand for interpreters, tools that decode the influence of input features on a DNN's decisions, especially in critical areas like healthcare and autonomous vehicles. Effective explanations enhances user trust, highlight model biases and also its strengths, fostering wider acceptance of these systems [3, 19, 34].

Within this field, perturbation-based methods are those which attempt to explain the machine learning model by connecting input modifications with output changes to construct an explanation heatmap, *i.e.*, a 2D attribution matrix indicating the responsibility of each input pixel to the model prediction [8, 15–17]. In that sense, occlusion is one of such methods, measuring the responsibility of each pixel by replacing image regions with a given baseline, *e.g.*, setting it to zero, and measuring output variations [28, 41]. Nevertheless, careless occlusion likely generates images which are outside of the training data's distribution, leading to unfair comparisons and fragile visualizations [19].

In order to address this shortcoming, we propose a novel interpretability framework that integrates naïve occlusion with other common image augmentations employed during model training. Our proposal hinges on a simple premise: if data augmentations are pivotal in model training, they can be equally instrumental in enhancing interpretability as the model's reaction to augmentations is a viable path to understand its decision-making process. However, seamlessly integrating these augmentations is not trivial. For example, *if jittering the color of an image changes the model output, how to pinpoint which region was most affected by it?*

Thus, a challenge arises when trying to determine the specific impact of an augmentation. Our approach relies on the DNN deep feature vector from the final layer before the classification head. We feed both the original images and their augmented variants (with or without occlusion) to CNNs or Vision Transformers. This yields two sets of deep feature vectors: one from original/augmented images without occlusion and another from their occluded counterparts, as depicted in Fig. 1.

We then compactly represent each set as a low-dimensional subspace in the deep feature vector space by applying Principal Component Analysis (PCA) without data centering to the set. Two subspaces $\mathcal{V}_M$ and $\mathcal{V}$ are generated from the sets extracted from images with and without occlusion, respectively. The core idea of our proposal is to measure the small perturbation due to the occlusion by the structural similarity Sim between $\mathcal{V}_M$ and $\mathcal{V}$, which is defined using the multiple canonical angles $\{\theta\}$ between the subspaces [14]. A larger subspace distance (orthogonal degree), $1 - Sim$, signifies that the occluded region is cru-

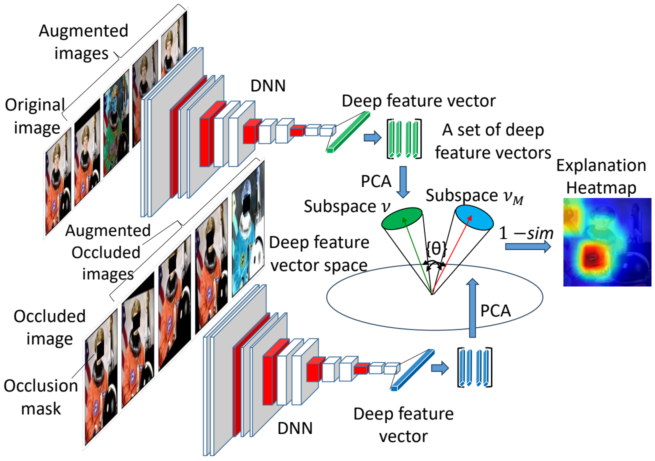


Figure 1. OSA-DAS Overview: Subspace $\mathcal{V}$ is derived from the augmented input image, while $\mathcal{V}_{\mathcal{M}}$ originates from its occluded counterpart. Both are derived from the principal component analysis (PCA) of a DNN's deep feature vector. The orthogonal degree [13, 14] between $\mathcal{V}$ and $\mathcal{V}_{\mathcal{M}}$ quantifies the occlusion's effect and shapes the explanation heatmap. Multiple occlusion augmentation subspaces are used to capture diverse facets of the input's representation. Their combined relationships offer a holistic view of occlusion impacts, producing a detailed heatmap.

cial for classification. This subspace representation method streamlines the process of merging multiple augmentation influences, offering a straightforward and robust metric of structural difference in the deep feature vector space.

Overall, our contributions are as follows:

1. We introduce a novel interpretability framework able to leverage any data augmentation to improve DNNs prediction explanation, shown in Fig. 1.

2. We leverage subspace representations [14] with the deep feature vector in explanation methods. This approach facilitates a more granular understanding of the model's behavior and offers a robust explanation.

3. We optimize our algorithm by designing a better random masking routine, which proposes better occlusions, allowing for a faster convergence.

4. We present a new interpretability metric named minimal size, which relies on causality theory [17] to measure how close the explanation heatmap is to the actual cause of the model prediction.

2. Related Work

The visualization of deep learning models decision-making process has become a vital research area, given the complex and often opaque nature of neural networks. Many methods have been introduced to shed light on how DNNs arrive at specific predictions. Gradient-based methods generate visualizations from the model output derivative with respect to the input image [31, 33]. Activation-based methods [6, 25, 26, 30, 32] build upon gradients but take into consideration common properties of the network structure, which improves output. These techniques can compute heatmaps quite fast yet many times lack explainability, showing many similarities to an edge detector [3, 20].

Additionally, given the recent developments in transformers [10, 22, 23, 38], a new family of attention-based interpreters has been proposed [1, 7], in which the attention weights from multiple layers are used to compute explanations. These methods demonstrate elevated interpretability capacity, but they are architecture-specific.

On the other hand, perturbation-based methods make minimal assumptions about the nature of the model itself and exactly for that reason show increased ability in explaining any kind of machine learning model. The basic perturbation method, Occlusion Sensitivity Analysis (OSA) [41], is actually quite straightforward. First, it measures the slight variation of the class score to occlusion in different regions of an input image using small perturbations of the image. Then, the resultant variation of each region is summarized as a heatmap of the input image. Other methods propose extensions to this idea by introducing new ways to generate the optimal occlusions [11, 12, 28, 36] or on how to compute their contributions [8].

Nevertheless, these methods are unable to explain the whole range of possibilities that can lead to a prediction, and have been criticized for analyzing the model on a different data distribution on which it was trained [19]. In that sense, the robustness of visual explanations to common data augmentation techniques, such as occlusions, has been studied. [35] analyzed the response of post-hoc visual explanations to natural data transformations. They found significant differences in robustness depending on the type of transformation, with some techniques demonstrating more stability. Similarly, [39] explored the relationship between data augmentation strategies and model interpretability, revealing that models trained with mixed sample data augmentation showed lower interpretability, particularly with CutMix [40] and SaliencyMix [37] augmentations. Moreover, [4] proposes an augmentation method leveraging multiple interpreters, thereby enhancing model robustness against noise or occlusions. This highlights the complex relationship between augmentation techniques and interpretability, raising caution for their adoption in critical applications. However, it's noteworthy that while these works analyze the impact of augmentations on explanations, as far as we know, none proposes an interpreter that leverages augmentation specifically to improve explanation trustworthiness.

3. Methods

In this section, we introduce our original method and metric. Details can be found in the supplementary material.

3.1. Occlusion with Augmentation Subspaces

Traditional occlusion sensitivity analysis (OSA) computes explanation heatmaps by replacing image regions with a given baseline (masking it to 0), and measuring the score difference in the output [28,41]. While this technique is cost-effective, occluded images originate from a distinct distribution from the one which the model was trained on. Thus, discerning whether the performance dip arises from this distributional shift or due to the responsibility of the occluded regions becomes ambiguous.

On the other hand, data augmentation (including random occlusions) have been used in most state-of-the-art models during training [5, 9, 27]. Therefore, we expect a more accurate interpretation could be performed if the model uses augmentations closer to the real training distribution.

With that in mind, we devise a technique that adapts OSA to using any data augmentation routine in an independent way by leveraging subspaces of deep feature vectors.

3.1.1 Data Augmentation Methods

Occlusion Sensitivity Analysis with Deep Feature Augmentation Subspace (OSA-DAS) utilizes data augmentation methods to foster more distinctive deep feature vectors that can be leveraged for enhanced interpretability.

In the realm of data augmentation, there exist prominent state-of-the-art routines that have revolutionized the process. For instance, RandAugment [9] is an automated data augmentation approach that streamlines the selection of transformations through two hyperparameters: n_{ops}, denoting the number of sequential augmentation transformations, and mag, representing the magnitude of these transformations. The transformations span from simple affine transformations, such as rotation and translation to more intricate operations such as color jittering and auto contrast.

On the other hand, TrivialAugment [27] presents an elegant yet powerful approach to automatic augmentation. It stands out due to its simplicity, requiring no parameters and applying a singular augmentation to each image. Despite its minimalist design, it has demonstrated its prowess, outperforming more complex augmentation techniques.

Central to our method is its adaptability and versatility. We chose the aforementioned augmentations in our experiments given they represent the pinnacle of current techniques, but our proposed framework is inherently flexible. It is designed to seamlessly integrate with any data augmentation routine, be it RandAugment [9], TrivialAugment [27] or else that best fits the explanation goal of the task at hand.

3.1.2 Deep Feature Augmentation Subspace

The addition of any data augmentation to perturbation-based interpretability is not trivial, and we opt to use sets of augmented inputs around each occlusion.

Consider that an image $\mathbf{x}$ is fed into the model $f\left(\right)$. In this paper, the output $\mathbf{v} = f\left(\mathbf{x}\right)$ is referred to as a deep feature vector in a k-dimensional vector space. For each occlusion $\mathbf{M}$, we generate a set of deep feature vectors corresponding to augmented images with occlusions, and then represent compactly the set by a subspace $\mathcal{V}_{\mathbf{M}} \subset \mathbb{R}^k$ for the specific occlusion. The same is performed for the original input image, which builds the reference subspace $\mathcal{V} \subset \mathbb{R}^k$.

The orthonormal basis, $\mathbf{V}$ and $\mathbf{V_M} \in \mathbb{R}^{k\times d}$, of the d-dimensional subspaces $\mathcal{V}$ and $\mathcal{V}_{\mathbf{M}}$ are calculated by applying Principal Component Analysis (PCA) without data centering to each set of deep feature vectors. More concretely, they can be obtained as the eigenvectors corresponding to several largest eigenvalues of auto-correlation matrix $\sum_{i=1}^{m} \mathbf{v}_i \mathbf{v}_i^T \in \mathbb{R}^{k\times k}$, where m is the number of applied augmentation types.

3.1.3 Structural Similarity between Two Subspaces

The relationship between two d-dimensional subspaces in $\mathbb{R}^k$ is defined by a set of d canonical angles $\{\theta_i\}_{i=1}^d$ between them. They can be obtained by applying singular value decomposition (SVD) to $\mathbf{V}^T\mathbf{V_M}$, where $\mathbf{V}$ and $\mathbf{V_M} \in \mathbb{R}^{k\times d}$ are the orthonormal basis [14]. The $\cos\theta_i$ of the i-th smallest canonical angle θ_i is the i-th largest singular value:

$$\cos\theta_i = \sigma_i\left(\mathbf{V}^T\mathbf{V_M}\right), \tag{1}$$

where $\sigma_i\left(\cdot\right)$ returns the matrix i-th largest singular value.

The structural similarity between two subspaces is defined as the sum of the square of the cosines of the first n_c canonical angles, where n_c is a hyperparameter indicating how much information from each subspace is to be considered [13, 14]. However, in our method, we need a measurement of subspace distance, which can be used as a proxy for the degree of responsibility r of each occlusion. Thus, we introduce the subspace distance, i.e., orthogonal degree, [14] defined by the following equation:

$$r\left(\mathbf{M}\right) = 1 - \sum_{i}^{n_c}\left(\sigma_i\left(\mathbf{V}^T\mathbf{V_M}\right)\right)^2. \tag{2}$$

3.1.4 Speedup by improved masking

OSA-DAS enhances OSA by incorporating more information, albeit at a higher computational cost. Essentially, perturbation-based interpretability is akin to a Monte Carlo approach for estimating machine learning models. The efficiency of this method can be improved by proposing better

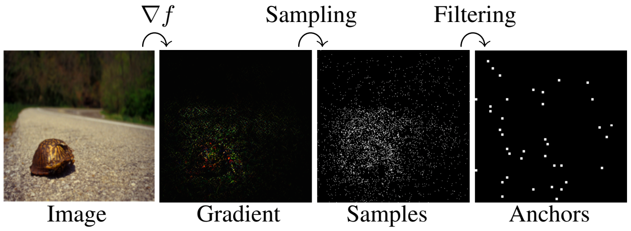


Figure 2. Mask anchor point selection via gradient sampling. The image gradient is produced on inference time, which is then used to sample anchor points. Anchor points too close to each other are filtered out. (anchors size is increased for visibility)

masks, thereby reducing the number of required masks, as seen in [36]. One straightforward strategy to devise superior masks is to utilize the model's gradient concerning the input image as weights. Albeit the simplicity, gradients are know to be noisy and not always indicate the most relevant features [32, 33], yet can be leveraged to sample the mask anchor points using a multinomial distribution. However, direct sampling often results in highly overlapping masks. To address this, we filter out those with substantial overlapping mask areas, as illustrated in Fig. 2.

3.1.5 Algorithmic generation of Explanation heatmaps

The presented ideas for the basis of our method is fully presented in Algorithm 1. Our OSA-DAS begins by sampling a set of augmentations on the original image. It then constructs a subspace, $\mathcal{V}$, which captures the model outputs for these augmented images. For each occlusion applied to the image, a similar subspace, $\mathcal{V}_\mathbf{M}$, is formed. The goal is then to compare the two subspaces, $\mathcal{V}$ and $\mathcal{V}_\mathbf{M}$, to understand the significance of the occluded region.

1. **Initialization:** Let f be a deep learning model that outputs a k-dimensional deep feature vector extracted from an input image. Let $\mathbf{x}$ be an input image and $\{\tau_i(\mathbf{x})\}_{i=1}^{n_a}$ a set of its augmentations. Besides, set parameters for the number of masks n_m, augmentations n_a, number of canonical angles n_c, and mask size l.

2. **Construct the Reference Subspace $\mathcal{V}$:** For i-th augmentation τ_i:

   (a) Feed augmented image $\tau_i(\mathbf{x})$ into the model f.

   (b) Normalize the length and store the deep feature vector $f\left(\tau_i(\mathbf{x})\right) \in \mathbb{R}^k$ in an array.

   We conduct the above process over all the augmentations, and then compute the orthonormal basis $\mathbf{V} \in \mathbb{R}^{k\times d}$ of the $\mathcal{V}$ subspace from the set of deep feature vectors $\{f\left(\tau_i(\mathbf{x})\right)\}_{i=1}^{n_a}$.

3. **Sample Masks and Construct Occluded Subspaces:** For each mask generated:

   (a) Create occlusions in the image using the mask.

   (b) For each occlusion, compute a basis $\mathbf{V}_\mathbf{M} \in \mathbb{R}^{k\times d}$ of subspace $\mathcal{V}_\mathbf{M}$ from the set of the k-dimensional feature vectors, $\left\{f\left(\tau_i\left(\mathbf{x}^M\right)\right)\right\}_{i=1}^{n_a}$, following the process in Step 2.

4. **Compute the orthogonal degree:** Measure the orthogonal degree $r(\mathbf{M})$ between the d-dimensional reference subspace $\mathcal{V}$ and occluded subspace $\mathcal{V}_\mathbf{M}$.

5. **Generate Explanation Heatmap:** Assign $r(\mathbf{M})$ to represent the significance of the occluded region $\mathbf{M}$. Combine these values across all occlusions $\{\mathbf{M}_i\}_{i=1}^{n_m}$ to form the heatmap $\mathbf{H}$. Normalize the heatmap to ensure values between 0 and 1.

Algorithm 1 Occlusion Sensitivity Analysis with Deep Feature Augmentation Subspace (OSA-DAS)

```
Require: $\mathbf{x} \leftarrow$ image, $f \leftarrow$model, $\tau_i \leftarrow i$-th augmentation
  $n_m \leftarrow$ number of masks
  $n_a \leftarrow$ number of augmentations
  $n_c \leftarrow$ number of canonical angles
  $l \leftarrow$ mask size
  $\mathcal{V} \leftarrow \{\}$
  for $i \leftarrow 1$ to $n_a$ do
      $\mathbf{x}_t \leftarrow \tau_i(\mathbf{x})$
      Insert the normalized $f(\mathbf{x}_t) \in \mathbb{R}^k$ in $\mathcal{V}$
  end for
  $\mathbf{V} \leftarrow PCA(\mathcal{V})$
  $\mathbf{H} \leftarrow 0$
  for $i \leftarrow 1$ to $n_m$ do
      $\mathbf{M} \leftarrow$ mask $(i, \mathbf{x}.shape, l)$
      $\mathcal{V}_\mathbf{M} \leftarrow \{\}$
      $\mathbf{x}^\mathbf{M} \leftarrow \mathbf{x} \odot \mathbf{M}$
      for $j \leftarrow 1$ to $n_a$ do
          $\mathbf{x}_t^\mathbf{M} \leftarrow \tau_j\left(\mathbf{x}^\mathbf{M}\right)$
          Insert the normalized $f\left(\mathbf{x}_t^\mathbf{M}\right) \in \mathbb{R}^k$ in $\mathcal{V}_\mathbf{M}$
      end for
      $\mathbf{V}_\mathbf{M} \leftarrow PCA(\mathcal{V}_\mathbf{M})$
      $r \leftarrow 1 - \sum_k^{n_c}\left(\sigma_k\left(\mathbf{V}^T\mathbf{V}_\mathbf{M}\right)\right)^2$
      $\mathbf{H} \leftarrow \mathbf{H} + (1 - \mathbf{M})\, r$
  end for
return $\frac{\mathbf{H}}{\sum \mathbf{H}}$
```

3.2. Explanation and Metrics

Even though the interpretability goal is to build clear visualizations of the machine learning model decision-making process, the comparison of interpreters at scale requires the application of metrics that can accurately measure the quality of the explanations [2].

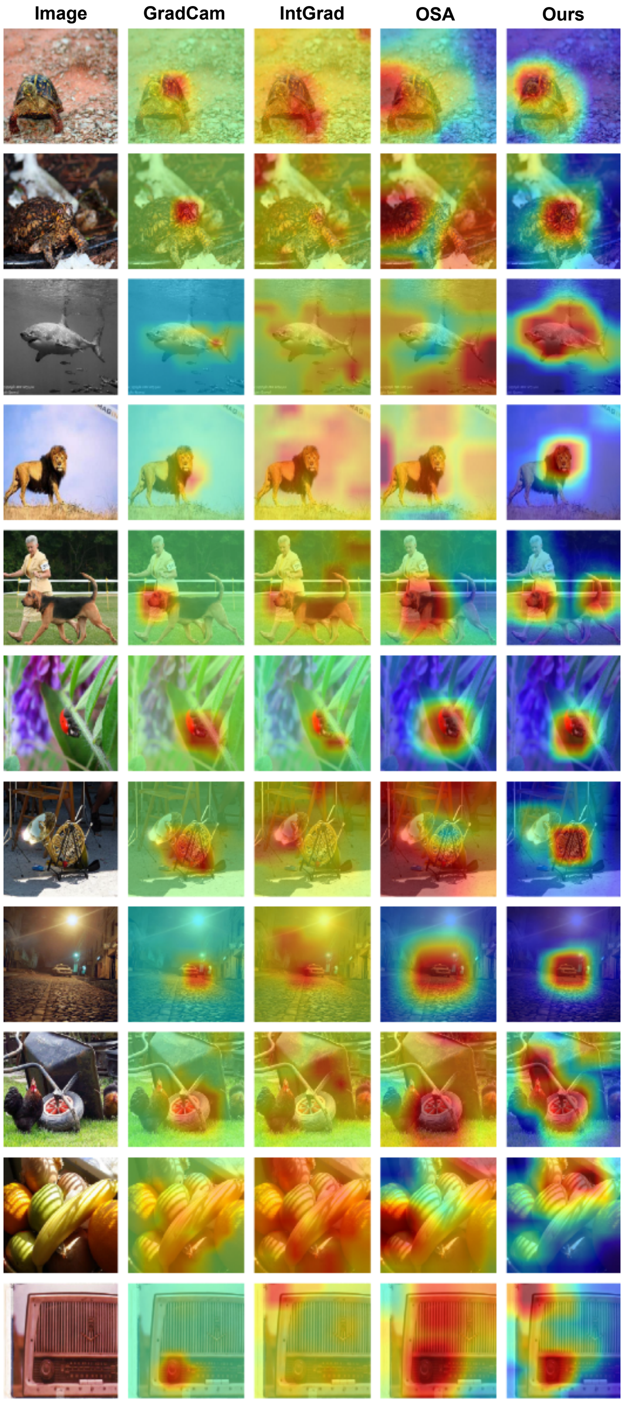


Figure 3. Explanation heatmaps visualizations for ResNet-50. Regions in red indicate the prediction causes. The proposed method generate concise and smooth explanation heatmaps, more in line to the general features the model is attending than other techniques.

3.2.1 Explanations

Given an input image $\mathbf{x}$, $\mathbf{S} = \mathbf{x} \odot \mathbf{M}$ indicates a masked subset of the input, where $\mathbf{M}$ is a binary mask and $\odot$ is the Hadamard product. Then, the explanation $\mathbf{E}$ is the minimal subset which has the same output as the original input.

$$\mathbf{E}\left(f|\mathbf{x}\right) = \min_{|\mathbf{S}|} \mathbf{S} : f\left(\mathbf{S}\right) = f\left(\mathbf{x}\right), \text{ with } |\mathbf{S}| > 0, \quad (3)$$

where $|\cdot|$ counts the number of unmasked pixels.

Eq. (3) is a general definition, and the nature of the model's output can vary depending on the algorithm. In this work, we want to build a class-agnostic method using deep feature vectors $\in \mathbb{R}^k$, which are extracted from the final layer before the classification head.

However, to compute the precise explanation using only Eq. (3) would require testing all possible subsets of pixels to ensure we have the minimal one [8]. In that sense, real interpreters provide an approximate explanation heatmap $\tilde{\mathbf{E}}\left(f|\mathbf{x}\right)$. This map is usually taken to be a description on how the model's predictions are influenced by each pixel [8, 17, 33]. In this work, we interpret these explanation heatmaps as probability distributions: they indicate the probability of each pixel in $\mathbf{x}$ belonging to the ideal explanation $\mathbf{E}\left(f|\mathbf{x}\right)$. See the supplementary material for details.

3.2.2 Evaluation Metrics

Many metrics have been proposed in interpretability literature, each offering different perspectives. In this paper, we chose to use multiple metrics to provide a more comprehensive measurement of the interpreter effectiveness. Deletion and insertion metrics [28] gauge the faithfulness of an explanation heatmap in representing a model's inferences.

First, the deletion metric measures how rapidly the model's prediction probability decreases when pixels are deleted according to their heatmap significance.

Conversely, the insertion metric evaluates how quickly the model's prediction probability escalates when pixels are inserted based on their heatmap significance. The performance of these metrics is quantified using the area under the curve (AUC), with the horizontal axis indicating the percentage of pixels deleted or inserted and the vertical axis representing the output probability of the model [7].

Although useful, we argue these metrics do not fundamentally align with the causality definition of explanation as per Eq. (3). Also, their numbers are not so intuitive and most often than not it is difficult to link their values to any visible property of the heatmap.

3.2.3 Minimal Size Metric

Sec. 3.2.1 defines an explanation as the smallest set of pixels that still results in the same model output. Now, given an explanation heatmap $\tilde{\mathbf{E}}\left(f|\mathbf{x}\right)$, we can try to generate an explanation from it. If the heatmap is correct, the explanation must use the minimal number of pixels, and we can use

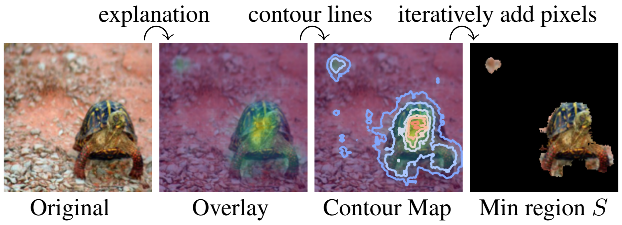


Figure 4. Simplified schema for computing the minimal size on a image-heatmap pair of 224×224 pixels with tolerance $\delta = 10^{-2}$. This simplified version decreases the iterations as follows: First, we divide the explanation heatmap into regions with the same importance level according to a contour map. Then, we introduce pixels from each region to the partial image in descending order of importance. We stop when the model's output of this partial image becomes very close to the one of the original image. The fraction of filled pixels in the partial image is the minimal size metric.

this number as a viable metric of the proximity between the explanation heatmap and the model's output cause.

$$s_{min}\left(\tilde{\mathbf{E}}\left(f|\mathbf{x}\right)\right) = \frac{|\mathbf{S}|}{|\mathbf{x}|}, \text{ with } f\left(\mathbf{S}\right) \approx f\left(\mathbf{x}\right), \tag{4}$$

where $f\left(\mathbf{S}\right) \approx f\left(\mathbf{x}\right)$ replaces the ideal equality $f\left(\mathbf{S}\right) = f\left(\mathbf{x}\right)$ in Eq. (3) to make the metric less rigid while also improving numerical stability.

We stress that our metric is class-agnostic, which allows us to directly use deep feature vectors $f\left(\mathbf{S}\right) \in \mathbb{R}^k$ and $f\left(\mathbf{x}\right) \in \mathbb{R}^k$, while the deletion and insertion metrics are exclusively based on the change of the scalar class probability change measured with AUC.

In practical terms, to compute this number we start from an empty set $\mathbf{S}$ and sequentially add pixels by order of importance, where pixel importance comes from $\tilde{\mathbf{E}}$. During this process, we must reach a point such that $||f\left(\mathbf{S}\right) - f\left(\mathbf{x}\right)||_1 \leq \delta$, where $||\cdot||_1 \leq \delta$ is an element-wise comparison within a fixed tolerance δ. Then, the algorithm stops and the ratio $\frac{|\mathbf{S}|}{|\mathbf{x}|}$ is returned.

Although this works, the number of steps can be reduced by adding batches of pixels instead of one pixel at a time, as exemplified at Fig. 4. Each batch is given from the contour map of $\tilde{\mathbf{E}}$, which splits the heatmap into regions by the intensity of each pixel, and determines the number of pixels to be added at each step.

Beyond that, notice a good interpreter metric should focus on evaluating only the explanation quality independently of model performance. This metric assesses the explanation's precision without being swayed by the model's accuracy while also providing a number that directly reflects the visual characteristics of the explanation. It's a clear and effective way to compare different interpreters' quality. See the supplementary material for more information.

3.2.4 Overall performance metric

While the Minimal Size metric offers a fresh perspective, it is essential to view it in conjunction with the currently used metrics for a holistic understanding. A more pivotal metric should thus be defined by balancing deletion, insertion and minimal size. We propose an Overall performance metric, building upon the work of [42], which combines insertion, deletion, and minimal size for a comprehensive evaluation.

$$overall = \frac{insertion - deletion}{minimal\ size} \tag{5}$$

Equation (5) offers a more thorough understanding of interpreter performance. The incorporation of size in the denominator ensures a dimensionless metric, where both the numerator and denominator represent areas. This combined metric offers a balanced and insightful evaluation of the general interpreter performance, making it a more sensible evaluator of the general interpreter performance.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed methods through comparison with the conventional explanation methods. This includes qualitative comparisons of explanation heatmaps and a quantitative evaluation using deletion, insertion [28] and minimal size metrics.

4.1. Experiment Settings

We employ ResNet-50 [18], ViT-B-14 [10] and Swin-V2 [22, 23] as classification models and assess the results on the validation set of ImageNet [29], which comprises 50K images from 1000 classes and is used in explainable AI literature for evaluation [7, 8, 28]. The images are resized to 256×256 pixels and center cropped to 224×224 pixels.

For our method, we use masks of $64 \times 64 (=l)$ pixels in the image as described in Sec. 3.1.4, TrivialAugment [27] as the augmentation routine, with $32 (=n_a)$ augmentations per occlusion. The dimensions of the deep feature vector was 786 for all models. For comparison, we perform the evaluations together with Guided Grad-CAM [30], Integrated Gradients [33] and OSA [41], which are frequently employed interpreters of each major family of methods. Given our emphasis on developing model-agnostic methods, we refrained from comparing with non model-agnostic methods, such as [1], [7], or expensive techniques like [24].

We used the implementation of these methods provided by the Captum tool [21]. The batch of all experiments performed in this work, including ablations, took approximately one week to run on 8 V100 16Gb GPUS.

4.2. Qualitative Results

The visualization results of the explanation heatmaps are showcased in Fig. 3 and Fig. 5. For the class-specific meth-

Table 1. Average Metric scores on ImageNet between ResNet-50, ViT-B and Swin-V2 models. For deletion and minimal size, lower is better (↓). For insertion and overall, higher is better (↑). **Bold** represents the best metric, while underline is the second best. Occlusion and Ours have the same mask size, but the former uses a sliding window, so it generates much more masks.

Method	**Minimal Size (↓)**	**Deletion (↓)**	**Insertion (↑)**	**Overall (↑)**
Guided Grad-CAM [30]	0.515	0.298	0.289	-0.034
Integrated Gradients [33]	0.518	**0.234**	0.267	0.123
Occlusion [41]	0.251	0.328	**0.549**	0.880
OSA-DAS (Ours)	**0.231**	0.331	0.539	**0.901**

ods, we show the heatmap with respect to the predicted class. These images illustrate how other interpreters tend to generate noisy heatmaps, especially notable for traditional OSA on model misclassifications, which attributes inverted responsibility compared to OSA-DAS.

Overall, the proposed method precisely captures the general features which the model attends to in a more stable manner, which facilitates model understanding and debugging. This resilience is likely due to its class-agnostic nature combined with the variety of feature comparisons enabled by the augmentation subspaces. These results suggest that OSA-DAS is capable of selecting the most impactful regions for the model, regardless of mispredictions.

On the flip side, the increased memory cost restricts the maximum number of masks and augmentations that can be applied, posing a trade-off for achieving more robust and accurate explanations.

4.3. Quantitative Results

Whereas Sec. 4.2 implies superiority of our proposed method, caution must be taken against sole reliance on visual assessments [3]. The average evaluation results on the whole validation set of ImageNet are presented at Tab. 1 using the metrics of insertion, deletion (Sec. 3.2.2), minimal size (Sec. 3.2.3) and overall (Sec. 3.2.4). Constant tolerance value $\delta = 10^{-2}$ is used. In deletion, the heatmap that accurately captures important individual pixels is highly valued, while for insertion, a heatmap presenting cohesive regions of importance is better evaluated [28, 36]. Minimal size metric measures proximity of the explanation to the actual cause. Overall balances insertion, deletion and minimal size areas evenly. This experiment uses 32 iterations for insertion, deletion and minimal size.

Integrated Gradients [33] and Grad-CAM [30] focus too much on important pixels, but not on important regions, which optimizes deletion in detriment of other metrics. Occlusion shows excellent insertion performance because it exclusively focuses on the regions which impact the predicted class. On the other hand, our method showcases best overall performance, showing good results among all metrics. We argue this demonstrates it can explain the actual prediction cause in a more holistic and class-agnostic way than others. Further details on the performance for each model is shown in the supplementary material.

In this context, it's noteworthy that class-specific methods, which consider specific priors for each class, are anticipated to perform better in insertion and deletion metrics compared to class-agnostic ones. This is because these methods priors (prediction probability) align with the same priors used in the evaluation metrics [7]. Regardless of such, our method still showcases comparable performance to OSA in spite of not being able to leverage such priors.

4.4. Ablation

We conducted an ablation study on our method's three key components: augmentations, masking, and subspace representations. These tests used 2500 ImageNet training images to ensure cost-efficiency and to maintain independence from analyses in Sec. 4.2 and Sec. 4.3. Tests used the ResNet-50 model, evaluating hyperparameters on a log2 scale until resource limits. We reported changes in the overall metric, defined in Sec. 3.2.4. Other metrics showed consistent behaviors and led to the same conclusions.

We examined how augmentations impact our method by switching from TrivialAugment [27] to RandAugment [9]. RandAugment allows for adjustable augmentation strength, even though the specific augmentations are random. Using 32 augmentations and 256 masks per image, we found, as seen in Fig. 6a, that our method remains stable up to a certain augmentation strength, beyond which it breaks down. This suggests the model can handle various augmentations as long as the image does not become unrecognizable.

Moreover, Fig. 6c demonstrates our technique possesses a better convergence rate with respect to the number of masks, with 256 masks already reaching good performance. This can be traced down to the efficient masking mechanism introduced at Sec. 3.1.4. In fact, the gradient is considered the simplest version of a gradient-based interpreter [31], and so, all we are doing is using a quick interpreter to derive a initial probability distribution for another interpreter. Thinking from the viewpoint of chaining interpreters, we can likely consider changing the gradient for other simple options, like Grad-CAM [30] for CNNs or Attention Rollout [1] for ViT [10]. Also, Fig. 6c shows

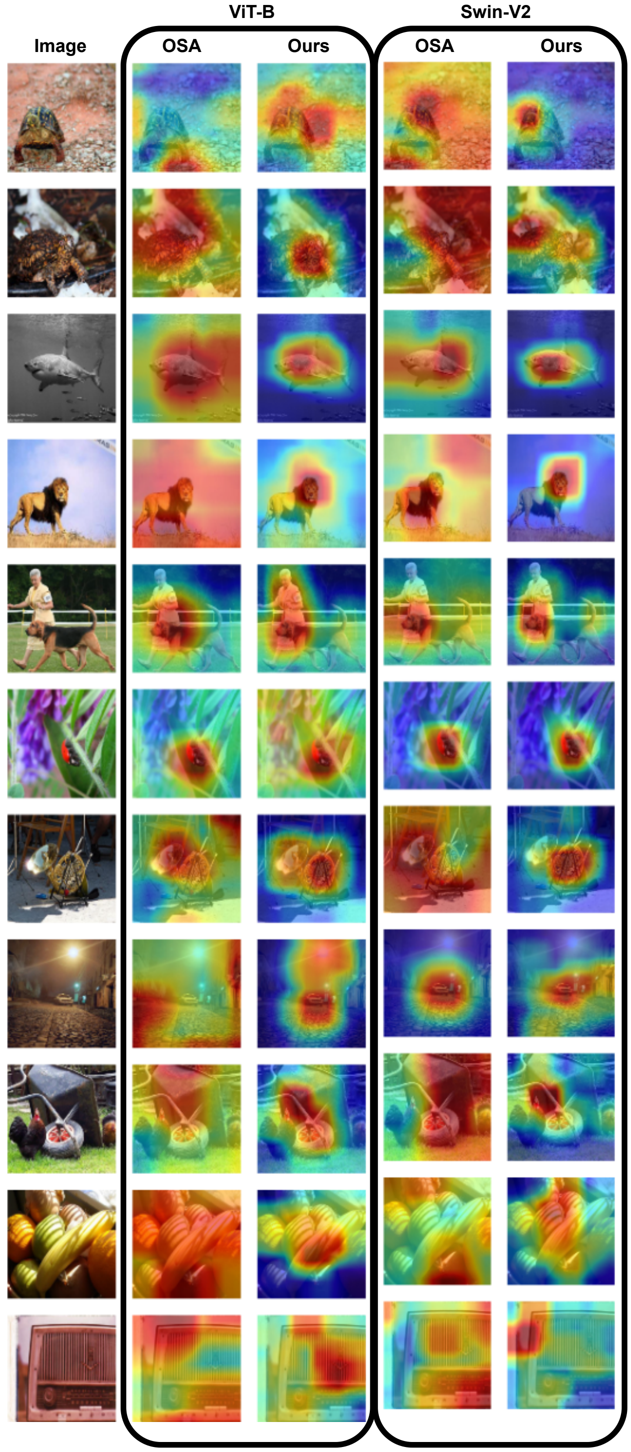
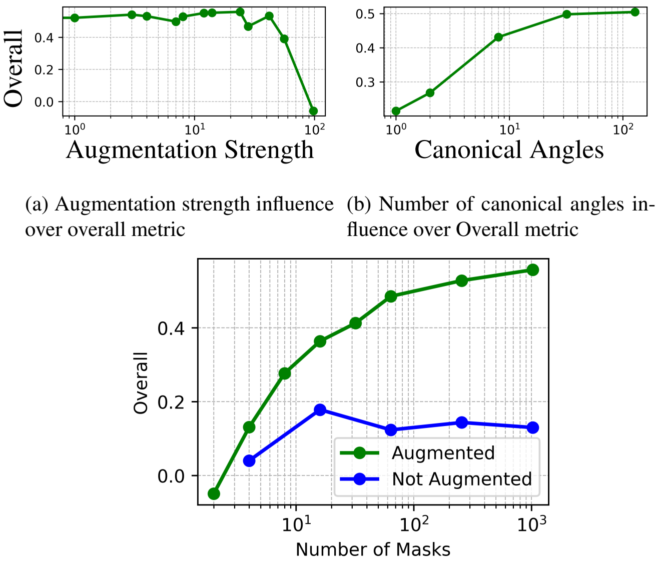


Figure 5. Explanation heatmaps visualizations for ViT-B and Swin-V2. Regions in red indicate the prediction causes.

that applying OSA-DAS without augmentations reduces its performance significantly, showing the convergence rate is correlated with the augmentations.



(a) Augmentation strength influence over overall metric

(b) Number of canonical angles influence over Overall metric

(c) Number of masks and augmentations influence over overall metric

Figure 6. Dependency of Overall metric with OSA-DAS hyperparameters. The horizontal axes are set to log scale for visibility.

Finally, we measure how many canonical angles should be used to measure the similarity between the original and occlusion subspaces. By this, we understand the impact of subspace representations to solve this problem. According to Fig. 6b, there is a clear dependency with n_c, but also not many angles are required to reach good performance. In fact, we can see the the curve starts to saturate after 32 angles (out of 786), which already provide over 2× improvement over using only 1 angle. We argue it is a strong favorable indicator for using subspace representations.

5. Conclusion

In this study, we proposed a model- and class-agnostic approach for interpreting machine learning model behavior based on general augmentations and occlusions, providing robust explanations for the decision-making process of computer vision models. Our contribution lies in the application of augmentations to occluded inputs and the use of subspace representation on deep feature vectors to gauge occlusion impact with improved precision. Moreover, we enhanced the computational efficiency by transitioning the occlusion selection process from random to gradient-based. Experimental results affirm our approach's superiority over traditional methods both quantitatively and qualitatively, providing sensible explanations that effectively demystify model decisions. This work heralds significant advancements in interpretability and trustworthiness of AI systems.

References

[1] Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers. In *ACL*, pages 4190–4197, Online, 2020. Association for Computational Linguistics. 2, 6, 7

[2] Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. Local Explanation Methods for Deep Neural Networks Lack Sensitivity to Parameter Values. In *ICLR*, 2018. 4

[3] Julius Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In *NIPS*, Oct. 2018. 1, 2, 7

[4] Flávio Arthur Oliveira Santos, Cleber Zanchettin, Leonardo Nogueira Matos, and Paulo Novais. On the Impact of Interpretability Methods in Active Image Augmentation Method. *Logic Journal of the IGPL*, 30:611–621, July 2022. 2

[5] Randall Balestriero, Ishan Misra, and Yann LeCun. A Data-Augmentation Is Worth A Thousand Samples: Exact Quantification From Analytical Augmented Sample Moments. In *NIPS*, Feb. 2022. 3

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *WACV*, pages 839–847, Mar. 2018. 2

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer Interpretability Beyond Attention Visualization. In *CVPR*, pages 782–791, June 2021. 2, 5, 6, 7

[8] Hana Chockler, Daniel Kroening, and Youcheng Sun. Explanations for Occluded Images. In *ICCV*, pages 1214–1223, Oct. 2021. 1, 2, 5, 6

[9] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, pages 3008–3017, June 2020. 3, 7

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, June 2021. 2, 6, 7

[11] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding Deep Networks via Extremal Perturbations and Smooth Masks. *ICCV*, pages 2950–2958, Oct. 2019. 2

[12] Ruth C. Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *ICCV*, pages 3449–3457, Oct. 2017. 2

[13] Kazuhiro Fukui. Subspace Methods. In *Computer Vision*, pages 1–5. Springer International Publishing, Cham, 2020. 2, 3

[14] Kazuhiro Fukui and Atsuto Maki. Difference Subspace and Its Generalization for Subspace-Based Methods. In *PAMI*, volume 37 of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2164–2177, 2015. 1, 2, 3

[15] Joseph Y. Halpern. A Modification of the Halpern-Pearl Definition of Causality. In *IJCAI*, International Joint Conference on Artificial Intelligence, May 2015. 1

[16] Joseph Y. Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach. Part I: Causes. In *The British Journal for the Philosophy of Science*, volume 56 of *The British Journal for the Philosophy of Science*, pages 843–887, 2005. 1

[17] Joseph Y. Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach. Part II: Explanations. In *The British Journal for the Philosophy of Science*, volume 56 of *The British Journal for the Philosophy of Science*, pages 889–911, Dec. 2005. 1, 2, 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, June 2016. 6

[19] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A Benchmark for Interpretability Methods in Deep Neural Networks. In *NIPS*, volume 32. Curran Associates, Inc., 2019. 1, 2

[20] Narine Kokhlikyan, Vivek Miglani, Bilal Alsallakh, Miguel Martin, and Orion Reblitz-Richardson. Investigating sanity checks for saliency maps with image and text classification. In *ICLR*, June 2021. 2

[21] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch. In *ICLR*, Sept. 2020. 6

[22] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. In *CVPR*, pages 11999–12009, June 2022. 2, 6

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *ICCV*, pages 9992–10002, Oct. 2021. 2, 6

[24] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *NIPS*, volume 30. Curran Associates, Inc., 2017. 6

[25] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. In *PR*, volume 65, pages 211–222, May 2017. 2

[26] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-CAM: Class Activation Map using Principal Components. In *IJCNN*, pages 1–7, July 2020. 2

[27] Samuel G. Muller and Frank Hutter. TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation. *ICCV*, pages 754–762, Oct. 2021. 3, 6, 7

[28] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *BMVC*, June 2018. 1, 2, 3, 5, 6, 7

[29] 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. In *IJCV*, volume 115 of *International Journal of Computer Vision*, pages 211–252, Dec. 2015. 6

[30] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128:336–359, Oct. 2016. 2, 6, 7

[31] K. Simonyan, A. Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR*, Dec. 2013. 2, 7

[32] D. Smilkov, Nikhil Thorat, Been Kim, F. Viégas, and M. Wattenberg. SmoothGrad: removing noise by adding noise, June 2017. 2, 4

[33] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *ICML*, volume 70, Sydney, Australia, June 2017. jmlr.org. 2, 4, 5, 6, 7

[34] Joe Biden The White House. President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, Oct. 2023. 1

[35] Lenka Tětková and Lars Kai Hansen. Robustness of Visual Explanations to Common Data Augmentation. In *CVPRW*, 2023. 2

[36] Tomoki Uchiyama, Naoya Sogi, Koichiro Niinuma, and Kazuhiro Fukui. Visually explaining 3D-CNN predictions for video classification with an adaptive occlusion sensitivity analysis. In *WACV*, pages 1513–1522, Jan. 2023. 2, 4, 7

[37] A. Uddin, M. Monira, Wheemyung Shin, TaeChoong Chung, and S. Bae. SaliencyMix: A Saliency Guided Data Augmentation Strategy for Better Regularization. In *ICLR*, June 2020. 2

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NIPS*, volume 30. Curran Associates, Inc., 2017. 2

[39] Soyoun Won, Sung-Ho Bae, and Seong Tae Kim. Analyzing Effects of Mixed Sample Data Augmentation on Model Interpretability, 2023. 2

[40] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *ICCV*, pages 6022–6031, Oct. 2019. 2

[41] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, volume 8689, pages 818–833, Cham, 2014. Springer International Publishing. 1, 2, 3, 6, 7

[42] Qing-Long Zhang, Lu Rao, and Yubin Yang. Group-CAM: Group Score-Weighted Visual Explanations for Deep Convolutional Networks, Mar. 2021. 6