IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 August 2025, accepted 7 October 2025, date of publication 11 November 2025, date of current version 24 November 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3631378

RESEARCH ARTICLE

IntActEval: A Continuous Framework for Quantitatively Evaluating Faithfulness of Feature Visualization Methods

BRADLEY GATHERS[1], IAN E. NIELSEN[1], (Member, IEEE), KEITH W. SOULES[1], OZAN TEKBEN[1], RAVI P. RAMACHANDRAN[1], (Senior Member, IEEE), NIDHAL CARLA BOUAYNAYA[1], (Member, IEEE), HASSAN M. FATHALLAH-SHAYKH[2,3], (Member, IEEE), AND GHULAM RASOOL[4], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ 08028, USA
[2]Departments of Neurology, The University of Alabama at Birmingham, Birmingham, AL 35294, USA
[3]Department of Mathematics, The University of Alabama at Birmingham, Birmingham, AL 35294, USA
[4]Machine Learning Department, Moffitt Cancer Center, Tampa, FL 08028, USA

Corresponding author: Ian E. Nielsen (nielseni6@rowan.edu)

This work was supported in part by the Army Research Office under Grant W15QKN-21-C-0077. The work of Ian E. Nielsen was supported by U.S. Department of Education through a Graduate Assistance in Areas of National Need (GAANN) Program Award under Grant P200A180055. The work of Ghulam Rasool was supported in part by NSF under Grant 2234468 and Grant 2234836.

ABSTRACT Trust is a critical factor in the safe and effective deployment of Artificial Intelligence (AI) models in essential tasks. With AI models increasingly being employed in domains such as self-driving cars, medicine, defense, and information technology, there is a pressing need to improve their explainability, trustworthiness, and interpretability. Feature Visualization (FV) is an approach that generates images to highlight the learned features of deep neural networks. However, numerous FV methods exist, and there is currently no standard framework to evaluate their effectiveness in improving model trust. This paper introduces a novel method, Integrating Activations to Evaluate Faithfulness (IntActEval), which quantitatively assesses FV methods by analyzing the faithfulness and accuracy of their visualizations to the model itself. We examined five FV methods across seven convolutional neural network (CNN) models. The Vanilla (unregularized) and Gaussian Noise (regularized) FV techniques produced the most faithful explanations for all seven models, with statistical significance for the tested data. In our CNN experiments, robustly trained models achieve the most plausible results. This paper provides a general guide to current FV methods and identifies the most reliable and effective techniques for enhancing the debugging and improvement of AI models.

INDEX TERMS Explainability, robustness, feature visualization, faithfulness.

I. INTRODUCTION AND MOTIVATION

Machine learning (ML) and artificial intelligence (AI) have positively influenced the global economy and have rapidly gained interest from companies, universities, healthcare institutions, government agencies, and the general public in the past decade. Its use in our daily lives is only accelerating with widespread applications that include but are not limited to biometrics [1], finance [2], online shopping [3], self-driving cars [4], healthcare [5], [6], autonomous drones to deliver shipments [7], social media [8], smart home devices [9], natural language processing [10], computer vision [11], military [12], transportation [13], and information technology [14]. These advancements have almost exclusively been through the use of Deep Neural Networks (DNN). The crucial question of how one can trust the decisions of these models often arises, especially for mission-critical tasks. The field of eXplainable Artificial Intelligence (XAI) seeks to elucidate

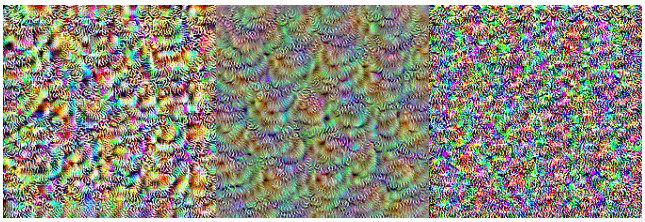
The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung.

how a model arrived at a decision, thereby enhancing public trust. There is a legal aspect to user trust in that the General Data Protection Regulation (GDPR) has been adopted by the European Union in May 2018 [15], [16]. Within this law, it outlines the need for an explanation provided to stakeholders for algorithmic decision-making systems that make decisions with legal effects when operating without human intervention [15], [16].

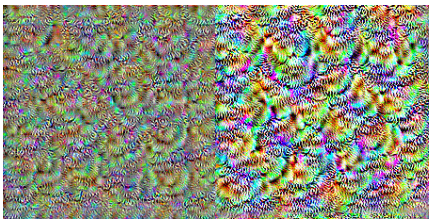
There have been numerous attempts to create visual explanations of what a machine learning model actually learns [17], [18], [19], [20]. When a machine learning model makes a decision, it takes into account information from all input features. However, some features contribute more to a model's decision than others (known as feature importance). One class of XAI methods, called attribution mapping, creates a map of importance scores attributed to each pixel in the input. This score reflects the significance of each pixel in generating a model decision. There have been some attempts to evaluate the faithfulness (measure of how precise the attribution map is to the model) of attribution methods [21], [22], [23], [24]. The Evaluating Attributions by Adding Incrementally (EvalAttAI) approach establishes consistency in quantitatively evaluating the faithfulness of an attribution map [22].

Another class of methods is based on feature visualization (FV). FV generates an image that depicts the learned abstract features that are vital in making a decision. Layer by layer, a DNN learns new and increasingly complex features that an FV will illustrate. FV commonly uses activation maximization (AM) that iteratively adjusts the input to maximize the activation of a neuron or layer of a DNN. This adjusted input provides a comprehension of the learned features. The problem of AM was first formulated in [25].

Figure 1 depicts FVs obtained using five different methods. In all five cases, the ResNet18 [26] convolutional neural network (CNN) with the same model weights was used. Also, the FV was taken at the last layer of the CNN. The Vanilla approach (see Fig. 1) is the simplest form of AM that uses only the gradient to maximize the activation of the neuron or layer of interest (referred to in the sequel as Vanilla AM). This is an example of an unregularized technique. The other four approaches illustrated in Fig. 1 include the use of a regularization term alongside the gradient which helps prevent the FV map from getting stuck in local minima. Frequency Penalization (FP) incorporates any regularization term that minimizes the variance among neighboring pixels of the FV [17], [27], [28] to maintain spatial continuity that reflects real-world data. Transformation Robustness (TR) techniques perturb the image slightly at each iteration prior to the forward pass through the model [27], [29], [30] such that a larger feature space can be explored. The TR and FP techniques are combined and referred to later as TRNFP. The Gaussian noise FV methods use only the gradient, but as a regularization step, add Gaussian noise to the input image. This is done prior to the forward pass at each step to simulate noisy data not seen during training.



(a) Vanilla (b) TRNFP (c) Gaussian Noise



(d) FP (e) TR

FIGURE 1. Comparison of FV methods on ResNet18 [26]. Each FV was generated via activation maximization (AM) from the final convolutional layer.

The trustworthiness of FV methods is questionable since the five approaches produce drastically different visualizations as shown in Fig. 1. This problem has been discussed in the literature as the disagreement problem [31], [32], where these differences in explanations are quantified. Since the same model was used for each method, the inconsistency in the results reveals that not all techniques are faithful. There have been some attempts to evaluate the faithfulness of attribution methods [21], [22], [23], [24]. Such a quantitative assessment is lacking for FVs and motivates our work.

In order to trust these models for critical tasks, we must also utilize robust machine learning to resist noise and adversarial attacks [33], [34], [35]. Compared to their non-robust counterparts, robust models maintain a relatively high accuracy on data that has a different distribution than what is seen during training. Robust models have also been shown to produce explanations that look more visually plausible to a human observer than their non-robust counterparts [36], [37]. This is significant because plausibility indicates that the model is learning the relevant features. However, it has been shown that these more visually plausible explanations generated by robust networks are not necessarily more faithful to the model itself [22]. For models and their explanations to be trusted, it is important to emphasize the need for visualizations to be both plausible and faithful. This will require robust models to produce more faithful visualizations, rather than varying the methods of FV and optimizing for plausibility alone.

In our prior work, EvalAttAI [22], we introduced a novel quantitative faithfulness metric for evaluating attribution maps. In this paper, we address a significant research challenge by proposing a new approach to assess the faithfulness of FVs. The novel contributions of this paper are as follows:

- We introduce a novel approach for quantitatively assessing the faithfulness of FVs and denote it as Integrating Activations to Evaluate Faithfulness (IntActEval).

- We conduct a comparative study to identify the most faithful FV method.
- We provide insights into the methods, model architectures, and training schemes that should be taken into account when attempting to explain a model.
- We investigate the potential correlation between the robustness of models and the faithfulness of their corresponding FVs.

The outline of this paper is as follows. Section II provides a concept map and overview of all explainability concepts and terminology relevant to our novel evaluation method. We also provide background on FV methods, their limitations, and how they relate to the broader ideas of XAI. Section III discusses the guiding principles for evaluating FV approaches. We will then detail the mathematics and methodology relating to the FV methods used and our IntActEval approach in Section IV. In Section V, we will discuss the results of our evaluation of multiple FV methods, providing a ranking of which explanations are most faithful. Section VI records the conclusions of the investigation.

II. BACKGROUND

To motivate the IntActEval method, we provide important contexts, including local and global explainability, FV methods, limitations of the approaches, qualitative and quantitative evaluations, and explainability notions. Figure 2 shows a concept map that connects these XAI concepts. All explainability concepts explored in this section are discussed primarily as applied to image classification models. In the following sections, the words ''model'', ''neural network'', and ''network'' are used interchangeably. In the field of XAI, the two main approaches for explaining a model are global and local explainability (see Fig. 2). Both of these concepts are covered, but the primary contribution of this work focuses on evaluating global explanations.

A. LOCAL EXPLAINABILITY AND ATTRIBUTION MAPPING

Local explainability creates a visualization that represents what the model is seeing for a specific input and output [36]. The most common approach for local explainability is attribution mapping. For local explanations, there are a handful of papers that introduce quantitative measures for evaluating their faithfulness [21], [22], [23], [24]. While our prior work (EvalAttAI) [22] focuses on evaluating the faithfulness of local explainability, this paper explores evaluating global explainability as discussed in the next subsection.

B. GLOBAL EXPLAINABILITY AND FEATURE VISUALIZATION

Feature visualization is an example of a global explainability technique that is used to create a visual representation of the learned features in a neural network, irrespective of any specific input. It does this by maximizing the activations of a single neuron, filter, or layer of a DNN model [17], [38]. This allows for the visualization of the features that these specific parts of the network are sensitive to. This procedure reveals: (1) what features and patterns the network has learned, (2) how the network processes information and classifies images, and (3) where the model is prone to error. The technique is to begin with an input image of random Gaussian noise and iteratively add the gradient such that a specific neuron or layer activation is maximized over multiple passes through the model [38]. As a result, all neurons in those influential layers or filters are maximally activated together. FV has also been used to find failure modes for image classification tasks [39].

There are four main categories of FV that include unregularized, regularized, learned prior, and dataset examples as shown in Fig. 2. The first three use AM. As mentioned earlier, the Vanilla AM is an unregularized approach that only uses the gradient. In contrast, the regularized approaches (Gaussian noise, FP and TR) also use AM but either apply a modification to the input image before each pass through the model, or include a regularization term that is added to the gradient. These approaches are visualized in Fig. 1. Learned prior methods introduce additional layers at the input of the network that are backpropagated when generating the FV. Finally, dataset example approaches choose samples from the data that highly activate the neuron(s) of interest to achieve a FV. Some popular implementations of these FV approaches include Deep Dream [29], Searching for High Activation Training Images [40], Priors and Generative Models [41], [42], and those that are AM based [43], [44].

C. EVALUATING EXPLAINABILITY: CRITERIA AND DEFINITIONS

There are a multitude of FV methods [25], [27], [29], [30], [38], [40], [41], [42], [43], [45], many of which produce vastly different results. While many of these methods have been researched, evaluating these explanations is a largely overlooked topic. In this section, we introduce the explainability criteria of robustness, plausibility, and most importantly for this paper, faithfulness. Our method, IntActEval, is a quantitative metric for assessing the faithfulness of a FV method.

1) ROBUSTNESS

As mentioned earlier, a robust model implies an acceptable classification accuracy on noisy data, adversarial examples, and data that has not been seen during training. However, robustness can have a negative impact on the accuracy of the model on clean data [46], [47]. Robustness can also have a significant effect on explainability. In fact, robust models have been shown to produce more visually plausible and interpretable results [22], [36], [37], [47]. Some models are designed with robustness in mind, resulting in an architecture that is inherently resilient to noise and attacks. One such

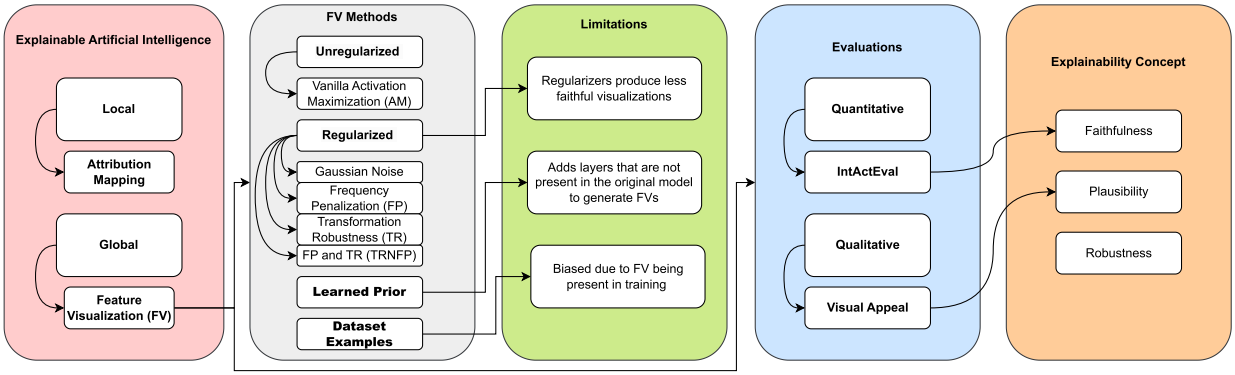


FIGURE 2. A visual representation of the connections between the various explainability concepts and techniques explored in this study.

model is PremiUm-CNN [33] which propagates uncertainty through each layer.

Achieving a robust model via training is the most common approach, as it can be applied to any existing model. Models can be robustly trained by adding noise to the input data before training. This noise can be, but is not limited to, Gaussian noise, salt and pepper noise, and adversarial attacks. In this paper, we used Gaussian noise, and the adversarial attack method of Projected Gradient Descent (PGD) [48] to robustly train our models.

2) PLAUSIBILITY

The plausibility of an explanation is most often a subjective criterion that describes how pertinent the explanation looks to a human observer. For instance, a highly plausible FV of a bird would contain many features that a human could recognize as pertaining to a bird. These may include patterns such as beaks, feathers, and wings. Generally, this criterion is only measured qualitatively. There are some quantitative metrics to evaluate the plausibility of attribution maps [49], [50]. However, a similar quantitative standard has yet to be established for FV. Studies have demonstrated that robustly trained models produce more plausible explanations [36], [37], [47], [51], [52]. This is a logical outcome, given that robust models learn features that enable them to correctly categorize data beyond their training distribution, thereby focusing on relevant features rather than high-frequency noise. Our results in this paper are consistent with this as well. It must be noted that plausibility and faithfulness do not appear to be correlated. However, it is important for explanations to be both plausible *and* faithful so that we can trust the explanations while also confirming that the model is recognizing relevant features. However, the focus of this paper is on faithfulness, as discussed next.

3) FAITHFULNESS

A faithful explanation is one that is highly accurate to the model itself. For an explanation to be useful for understanding and debugging, it must be faithful. In the context of FVs, this means that the explanation should truly maximize the activation for the neuron or layer of interest the most. Our novel approach, called IntActEval, uses this concept and provides a score and ranking for a range of FV methods.

D. LIMITATIONS

For this paper, we assess the limitations of FV methods through the lens of faithfulness.

1) REGULARIZATION

The unregularized method of FV (uses AM) and the regularized method based on adding Gaussian noise tend to produce less plausible explanations than other regularized and learned prior methods. However, this paper focuses on finding which methods are most faithful to the model itself. Therefore, less plausible explanations indicate that the learned features of the model are less salient and should not reflect negatively on the explainability methods. As we will discuss in the results, certain regularized methods produce less faithful explanations, at least after many iterations. This is due to the fact that these regularizers restrict the gradient, which in turn makes it harder for the model to optimize towards a faithful representation of the object specified. This impedes the explanations from producing larger activations beyond a certain scale of perturbation.

2) LEARNED PRIOR

The learned prior approach also introduces a major limitation, which is that it requires additional layers that are not present in the actual model during training or testing. Therefore, this approach creates a visualization for a slightly different model. This visualization might be faithful to that modified model, but not to the original model being explained.

3) DATASET EXAMPLES

Finally, using highly activating dataset examples also has the limitation of being unable to produce large activation scores when compared to other approaches, and thus they are not as

faithful. In addition, these examples do not help show how well the model will deal with out-of-distribution data, as the examples are from the distribution it was trained on.

E. RELATED WORK

The challenge of ensuring that explanations from eXplainable AI (XAI) methods are faithful to a model's underlying reasoning has spurred a significant body of research. However, the existing evaluation landscape is overwhelmingly dominated by methodologies designed for local explainability, which assess why a model made a specific decision for a single input [53], [54]. This includes popular techniques like DeConvNet [55], Class Activation Mapping (CAM) [50], and Gradient-weighted Class Activation Mapping (Grad-CAM) [56], which produce attribution maps highlighting important input regions. Consequently, the prevailing evaluation paradigms—including perturbation-based metrics like ROAR [21], axiomatic approaches [57], synthetic ground-truth frameworks [58], and sanity checks [59]—all share this focus on local, attribution-based explanations. While foundational, these methods are fundamentally designed to assess where in a given input the model focuses. This leaves a conspicuous gap, as there is no established quantitative framework for assessing the faithfulness of global explanations, such as Feature Visualization (FV), which aim to show what a model has learned in general, independent of any single input.

This lack of appropriate tooling has meant that the evaluation of global, generative FV methods often reverts to subjective assessments of visual plausibility [60], a criterion that can be misleading and has been shown to be uncorrelated with faithfulness [61]. Our work, IntActEval, directly addresses this critical research gap. Unlike the local attribution-focused methods, IntActEval provides the a novel quantitative, continuous, and objective framework specifically designed to evaluate global FV techniques. It assesses the core claim of any FV method: that the generated image maximally activates the model unit it purports to represent. By using the model's own activation outputs as the ground truth, IntActEval introduces a much-needed, objective benchmark for a class of explainability methods that has, until now, largely escaped rigorous faithfulness assessment.

III. FEATURE VISUALIZATION EVALUATION

There is currently a strong absence of quantitative methods for evaluating the faithfulness of FV techniques. Generally, papers on the topic evaluate the FV maps subjectively based on the perceived plausibility of the explanations [17], [38]. The absence of standardized criteria for evaluating faithfulness across these diverse methods currently limits confidence in their widespread deployment. This issue motivates our work.

We now introduce some key principles of FV faithfulness evaluations, informed by prior work in explainability assessment [22]. The IntActEval method is underpinned by a set of propositions that establish a benchmark for evaluating the faithfulness of an explanation. The widespread lack of ground-truth explanations and standardized evaluation protocols has created a landscape of heterogeneous and often contradictory evaluation results. The IntActEval method operationalizes the following propositions, which aim to address common pitfalls in explanation evaluation:

- **Self-Tested:** A credible faithfulness evaluation should leverage the information provided by the model itself to assess and score the explanation. This principle confronts the subjectivity and scalability crisis of human-centric evaluations [62]. Relying on human annotators is resource-intensive and slow, introduces bias and low reproducibility [63], and often confuses faithfulness with plausibility [64]. By using a functionally-grounded approach with the model's own outputs as the evaluation signal, the self-tested principle ensures objectivity and a direct link to the model's computational behavior rather than human perception.
- **Iterative:** Given that FVs generate explanations in an iterative manner, any corresponding assessment should also evaluate the visualization at each stage to capture the complete narrative. Evaluating an explanation at a single point in time can be misleading. The faithfulness of an explanation can change significantly during its generation, similar to how attribution method performance varies across different network layers [65]. A longitudinal evaluation design is necessary to provide a comprehensive understanding of the explanation's behavior over time.
- **Continuity:** To ensure continuity, all scores at each iteration should contribute to the final faithfulness score. IntActEval accomplishes this by calculating the area under the curve (AUC) of the activations over time. Simpler metrics like the final score or a simple average discard important information about the explanation's quality trajectory. The use of AUC is a well-established technique to summarize performance over an interval [66]. It provides a holistic measure that rewards methods for achieving and maintaining high faithfulness throughout the entire process. By maintaining continuity, the evaluation remains *intact* and is the inspiration for the name IntActEval.
- **Fairness:** In the interest of fairness, each FV method should be assessed using the same initial image (in our case, random noise). Furthermore, the perturbations at each step should be normalized to maintain the same mean and variance, preventing any FV method from gaining an unfair advantage due to more significant changes at each step. To ensure a fair comparison, all methods must start from an identical, information-neutral point to control for initialization effects [65]. Normalizing the update step ensures that performance differences are due to the quality of the method's guidance, not the magnitude of its perturbations.

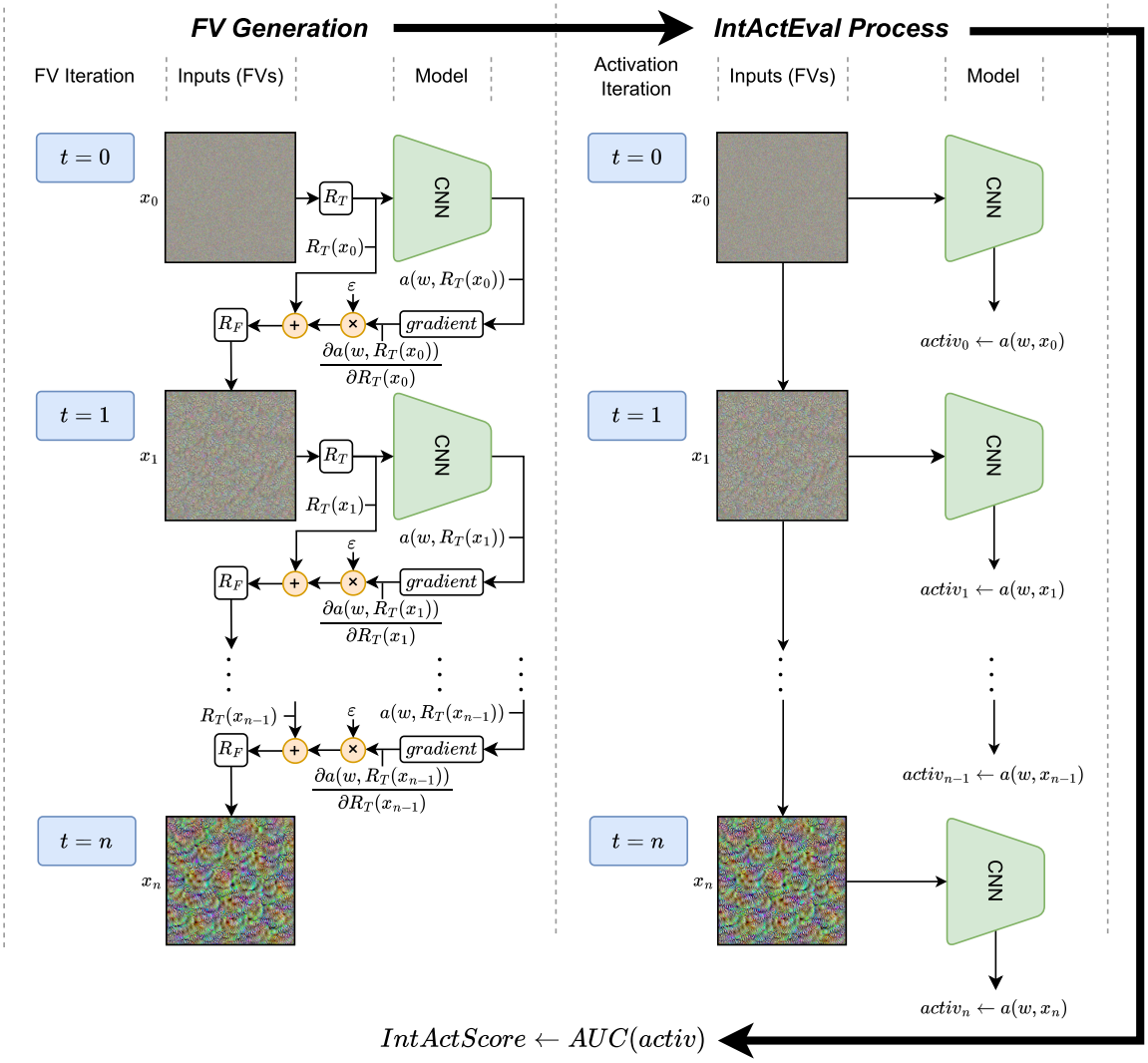


FIGURE 3. The process of generating FVs using TRNFP and evaluating the faithfulness of the FVs using IntActEval. The left panel illustrates FV generation, while the right panel details the IntActEval scoring process (see Pseudocode 1 for implementation). Identical evaluation steps are applied to all FV methods compared in this study: Vanilla (no regularizers), TR-only, FP-only, and Gaussian. Gradients are zeroed before each forward pass to prevent their accumulation and to isolate the computation for each pass.

IV. METHODOLOGY AND EXPERIMENTS

This section provides a comprehensive overview of the FV methods used, the novel IntActEval approach, the investigated models, the dataset utilized, the training methods employed, and the convergence and running time aspects of the iterative IntActEval method.

Figure 3 depicts (1) the FV generation and (2) the IntActEval method for the FV evaluation that is accomplished by inputting the FVs to the model. Although Fig. 3 depicts one particular FV generation method (TRNFP, which is a combination of TR and FP), the concept carries over to the other approaches. All FVs are generated prior to any evaluation by the model. However, this is also equivalent to evaluating the FVs after each iteration, thereby minimizing memory usage for explanations. For clarity of presentation, we describe the process as distinct stages of FV generation and evaluation.

A. FEATURE VISUALIZATION METHODS

As artificial intelligence becomes more integrated into our daily lives via mainstream technology, it is crucial to understand the underlying principles that govern them. FV can aid in this understanding. There are four main classes of FV approaches, as described earlier in Section II-B. The approaches explored in this study include Vanilla AM [25], Transformation Robustness (TR) [29], Frequency Penalization (FP) [27], combination of TR and FP (TRNFP), and adding Gaussian noise [41]. These approaches can be seen visually in Fig. 1.

The first step of each approach uses an input image x_0 of random zero-mean Gaussian noise (denoted as $N(0, \sigma)$ where $\sigma = 0.5$) to generate the first FV x_1. All subsequent steps use the FV generated in the previous iteration as the input. The TR and FP methods also use regularization. Equation 1 is a

general expression for updating the FV x_t at iteration t to the FV x_{t+1} at iteration $t+1$ as given by

$$x_{t+1} = f(x_t, \varepsilon * \frac{\partial a(w, x_t)}{\partial x_t}) \tag{1}$$

where f describes a function of x_t and the gradient ($\frac{\partial a(w,x_t)}{\partial x_t}$) which is the partial derivative of the activation function (denoted as $a(w, x_t)$) with respect to the model input. This function f is defined for each particular FV method in Eqs. 2 to 6. Note that w represents the model weights and ε is a small value that is multiplied by the gradient to control the step size of each iteration (ε is assigned the same value across all methodologies to ensure fairness). The gradient is calculated via backpropagation. This iterative process of getting x_{t+1} from x_t repeats until the desired number of FVs are generated.

The activation function ($a(w, x_t)$) is the output of the chosen node, layer, or channel that we are seeking to maximize. In our case, it is the mean of all activations that are output from the final convolutional layer of the model. This value is known as the activation score. This layer was selected due to its ability to yield distinct FVs which effectively represent a composite of high-level learned features when its activations are maximized for [17], [67]. The process of maximizing the activation function within the model results in the FV at each iteration being a better visual representation of the learned features compared to the preceding iteration.

In the experiments, each FV method was run for 250 (i.e. $n = 250$) iterations over 10 generations, such that all the methods were fairly evaluated. The Vanilla method gives a baseline with no regularizers. All FV methods were generated on both standard and robustly trained models as described later.

1) VANILLA AM

Vanilla AM is the simplest approach, and serves as the basis for all other FV methods used in this paper. Equation 2 depicts one iteration for the Vanilla AM FV method as

$$x_{t+1} = x_t + \varepsilon * \frac{\partial a(w, x_t)}{\partial x_t}. \tag{2}$$

When using Vanilla AM, the resulting visualizations often appear to contain more high-frequency noise than other more complex methods of FV. However, despite the more plausible looking explanations from more sophisticated approaches that use regularization to reduce noise, these methods might not be as faithful to the model. The subsequently described FV methods (FP, TR and Gaussian) all use regularization strategies, but only the first two reduce noise.

2) TRANSFORMATION ROBUSTNESS (TR)

The TR approach applies regularization to the image prior to the forward pass of the model. When using the TR approach, Eq. 2 is modified as given by Eq. 3:

$$x_{t+1} = R_T(x_t) + \varepsilon * \frac{\partial a(w, R_T(x_t))}{\partial R_T(x_t)}, \tag{3}$$

where R_T is the regularizer applied to the input FV x_t. This regularizer is only applied every few iterations to allow for the maximization to occur, as the transformations decrease the activation scores that we are seeking to maximize.

TR techniques apply a transformation to the image with approaches such as stochastic jitter, rotation, or image scaling [29]. Although the perturbations are applied before any optimization takes place, the aim is to maximize the specified activations. This has proven to result in noticeably more visually appealing and plausible FVs [29]. Stochastic jitter ($\pm$ 4 pixels) and rotation ($\pm$ 5 degrees) were used for the generation of the FVs in our experiments. The jitter was applied every 10 iterations, and the rotation was applied every 20 iterations.

3) FREQUENCY PENALIZATION (FP)

FP methods aim to reduce the variation of pixels in the image by implementing a variety of regularization techniques. Some commonly used regularizers include L2 decay [68], [69], bilateral blur [45], normalized and contribution crop [28], and total variation [27]. This in turn mitigates high-frequency noise in the generated visualizations. When using the FP approach, we apply the regularizer R_F after the optimization step as defined in Eq. 4 as

$$x_{t+1} = R_F(x_t + \varepsilon * \frac{\partial a(w, x_t)}{\partial x_t}). \tag{4}$$

Reducing the high-frequency noise also results in a more visually appealing and plausible FV [27], similar to TR. However, the resulting FV has been modified post-hoc for FP, rather than directly showing the input as maximized by the gradient. This disconnect could be the reason why FP creates more visually appealing explanations, but possibly at the cost of decreasing faithfulness.

In the experiments, the FVs were regularized with L2 decay (1E-5), bilateral blur (color and spatial blur at 0.5, kernel size of 3), normalized crop (15th percentile), and contribution crop (15th percentile). All FP regularizers were applied every iteration except for bilateral blur which was applied every 3rd iteration. Applying bilateral blur more often produces color artifacts on the FV. This was also seen if the color blur parameter was much higher than the spatial blur parameter.

4) TRANSFORMATION ROBUSTNESS AND FREQUENCY PENALIZATION (TRNFP)

We also apply both TR and FP together to generate FVs as depicted in Fig. 3. This is defined in Eq. 5 as

$$x_{i+1} = R_F(R_T(x_i) + \varepsilon * \frac{\partial a(w, R_T(x_i))}{\partial R_T(x_i)}), \tag{5}$$

where R_F and R_T denote the FP and TR regularizers respectively. In Eq. 5, the TR regularizer is applied prior to optimization and the FP regularizer is applied after optimization. The regularizers utilized align with the definitions provided for TR and FP earlier. The concurrent application

of both regularizers exerts a more significant influence on the final FV. This can enhance the visual appeal, but carries the potential risk of compromising the faithfulness due to the substantial modifications introduced both pre- and post-optimization.

5) GAUSSIAN NOISE

A regularization approach introduced in [41] adds a small amount of random Gaussian noise to the image during AM. These random perturbations allow for the visualizations to explore a larger feature space, thereby improving image diversity and avoiding local minima. However, this might lead to less plausible looking explanations due to the noisy character of these visualizations.

The Gaussian approach to FV is defined in Eq. 6 as

$$x_{t+1} = x_t + \varepsilon * \frac{\partial a(w, x_t)}{\partial x_t} + N(0, \sigma), \tag{6}$$

where the zero-mean Gaussian noise added to the input is represented as $N(0, \sigma)$ and σ signifies the variance of the distribution. The noise is generated with the same dimensions as the input x. This noise is applied at every iteration. For our experimentation, $\sigma = 0.5$. Notably, this regularization approach adds noise to the image rather than reducing it like TR or FP. Similar to Vanilla AM, the Gaussian method produces less plausible looking FVs than their less noisy alternatives, which could have the inverse effect on faithfulness.

B. INTEGRATING ACTIVATIONS TO EVALUATE FAITHFULNESS (INTACTEVAL)

We propose a novel method of evaluating the faithfulness of FVs quantitatively. It is an iterative process that stores and then evaluates the FVs generated at each step as described above and depicted in Fig. 3. The FVs are generated using Vanilla AM, TR, FP, TRNFP, and Gaussian Noise and then compared with respect to faithfulness.

Once FV generation has been completed, the model gradient is set to zero before the individual FVs are passed through the same network to obtain the activation functions ($a(w, x_t)$) (see Fig. 3). At each iteration number t, the resulting activation score (mean of all activations that are output from the final convolutional layer), $activ_t$, is obtained. The result is an activation score curve of $activ_t$ versus t. Next, the trapezoidal rule is used to calculate the area under the curve (AUC) of the activation score graph. This AUC denotes the faithfulness score or equivalently the IntActEval score. It is an approximation of the integral of the activations from which the name of our new quantitative evaluation method is derived.

Using the AUC is crucial because it provides a holistic measure of faithfulness, capturing the cumulative activation across the entire process rather than relying on a single, potentially misleading point like peak activation. A higher AUC signifies that the generated feature vectors consistently and strongly align with the model's internal representations, offering a more robust and comprehensive assessment of the explanation's faithfulness to the model's reasoning.

Pseudocode 1 FV and IntActEval Process

Initialization, Initial FV is random Gaussian noise
$x_0 \leftarrow N(0, \sigma), \sigma = 0.5$
$R_T(\cdot) \leftarrow regularizer(TR)$ ▷ TR Regularizer
$R_F(\cdot) \leftarrow regularizer(FP)$ ▷ FP Regularizer
Input number of iterations n
Generate FVs
for t in $0 : n$ **do**
Zero the gradient before each forward pass
    $model_{gradient} \leftarrow 0$
Vanilla
    $a(w, x_t) \leftarrow model(x_t)$
    $x_{t+1} \leftarrow x_t + \varepsilon * \frac{\partial a(w, x_t)}{\partial x_t}$
or TR
    $a(w, R_T(x_t)) \leftarrow model(R_T(x_t))$
    $x_{t+1} \leftarrow R_T(x_t) + \varepsilon * \frac{\partial a(w, R_T(x_t))}{\partial R_T(x_t)}$
or FP
    $a(w, x_t) \leftarrow model(x_t)$
    $x_{t+1} \leftarrow R_F(x_t + \varepsilon * \frac{\partial a(w, x_t)}{\partial x_t})$
or TRNFP
    $a(w, R_T(x_t)) \leftarrow model(R_T(x_t))$
    $x_{t+1} \leftarrow R_F(R_T(x_t) + \varepsilon * \frac{\partial a(w, R_T(x_t))}{\partial R_T(x_t)})$
or Gaussian
    $a(w, x_t) \leftarrow model(x_t)$
    $x_{t+1} \leftarrow x_t + \varepsilon * \frac{\partial a(w, x_t)}{\partial x_t} + N(0, \sigma = 0.5)$
end for return x_t for $t = 0 : n$
IntActEval Scoring
for t in $0 : n$ **do**
Zero the gradient before each forward pass
    $model_{gradient} \leftarrow 0$
    $a(w, x_t) \leftarrow model(x_t)$
Activation score at iteration t
    $activ_t \leftarrow a(w, x_t)$
end for return $activ$
Calculate IntActEval score
$IntActScore \leftarrow AUC(activ)$

Pseudocode 1 shows the experimental procedure used to generate and evaluate FVs. It depicts an algorithmic narrative of Fig. 3. The IntActEval method satisfies the four criteria established in Section III by being self-tested, iterative, continuous, and fair. Adherence to the self-tested principle is achieved by using the model as the ultimate arbiter of the explanation's quality. By propagating the generated FV from each iteration back through the network and measuring the target neuron's subsequent activation, IntActEval provides a functionally-grounded score.

The framework addresses the Iterative principle by assessing the FV at every step of its generative process, thus capturing the complete temporal narrative of the explanation rather than a potentially misleading static snapshot. This longitudinal data is then integrated into a final score with Continuity by calculating the Area Under the Curve (AUC) of the activation values over all iterations. This use of AUC provides a holistic and robust measure that rewards sustained faithfulness, effectively preventing the information loss that would occur with simpler metrics like the final activation value or a simple average.

To ensure a fair and standardized comparison, IntActEval enforces a level playing field. All FV methods are initialized from an identical starting point—a random noise image (x_0)—to control for initialization effects. Crucially, fairness is maintained throughout the process by normalizing the perturbation applied at each step, ensuring that no method gains an undue advantage from the magnitude of its updates. This protocol guarantees that observed performance differences are attributable to the quality of the guidance from the FV method itself, not to artifacts of the evaluation setup.

C. MODELS AND DATA

The seven models used in this paper include EfficientNet [70], ResNet18 [26], AlexNet [71], [72], DenseNet [73], GoogLeNet [74], SqueezeNet [75], and MobileNetV2 [76]. All of these models are widely used CNNs and come built-in with the Torchvision library [77]. We selected CNNs for evaluation due to (1) their widespread usage in computer vision and image classification, and (2) the availability of numerous FV methods specifically designed and tested using this architecture. We also decided to use CIFAR-10 [78] as the dataset of choice for training and evaluation. CIFAR-10 consists of 10 different classes and contains full color images that are 32 by 32 pixels in size. This dataset was chosen due to its relatively low resource requirements and its widespread use in CNN training.

D. TRAINING

In order to compare the faithfulness of FVs, seven different models were trained both standardly and robustly using the CIFAR-10 dataset. In standard training, the model is trained on clean data. In robust training, we applied adversarial noise to the data, using the same hyperparameters as the standard trained model. Adversarial attacks apply small perturbations to the input to trick the model into incorrect classification. Adversarial training was tested because this approach leads to a significant performance improvement for data outside of the training distribution [79], [80], [81]. Adversarial attacks were generated using the Projected Gradient Descent (PGD) method [79].

By training models with and without adversarial attacks, we are able to evaluate how model robustness affects FVs. We also train multiple models to account for differences in CNN model architectures.

For both standard and robust training, transfer learning [82] was implemented by starting with ImageNet-pretrained weights from the Torchvision library. All models were then trained on the CIFAR-10 dataset. This method of training was selected to reduce computational complexity. All layers were unlocked and trained during transfer learning, and the final output layer was altered to fit ten classes [83].

TABLE 1. Test accuracy (%) of the seven models using CIFAR10.

Model	Test Accuracy (%)	
	Standard	Robust
ResNet	95	92
AlexNet	85	82
DenseNet	91	89
EfficientNet	92	90
GoogleNet	91	89
MobileNetV2	91	89
SqueezeNet	86	84

The test accuracies of each of the seven trained models can be seen in Table 1. All 10,000 images from the test batch of the CIFAR-10 dataset were used.

E. CONVERGENCE AND TIME CONSIDERATIONS

The IntActEval method is iterative but there is no mathematical guarantee of convergence. We selected 250 iterations as the point to stop the algorithm. This was based on an empirical calculation that compares the activation score at iteration t ($activ_t$) with the activation score at iteration $t-1$. Specifically, Δ as defined in Eq. 7

$$\Delta = \frac{|activ_t - activ_{t-1}|}{activ_t} \tag{7}$$

was found to be less than 0.001 for all FV methods applied to MobileNetV2 after 250 iterations.

We investigate the time taken to run 250 iterations on our Quadro RTX 8000 GPU. Five FV methods were applied to both standard and robustly trained models. There are seven model architectures used in this study. This results in 70 time measurements. The number of time measurements is augmented to 700, since 10 trials were conducted for each FV method. The average time taken across these 700 time measurements is 7.90 seconds, indicating that the algorithms are both fast and accurate. Subjectively, FVs generated by each method at each iteration beyond 250 showed no observable difference for all the models used.

V. RESULTS AND DISCUSSION

This section presents the results of the faithfulness evaluation for multiple robust and non-robust models. The faithfulness score of each FV was analyzed using IntActEval, as described in Section IV. The results in this section can be broken down into three main parts: activation versus iterations plots, AUC (IntActEval score) of those activation curves, and a visual analysis of the FVs. The AUC bar plots show the final IntActEval scores for each method and model. We also discuss some FV usage recommendations and future work at the end of this section.

A. ACTIVATION RESULTS

In this subsection, we analyze the activation results and explain the calculation of the IntActEval score, which is derived from the AUC of the activation plots. During the IntActEval process outlined in Section IV-B, we generate FVs

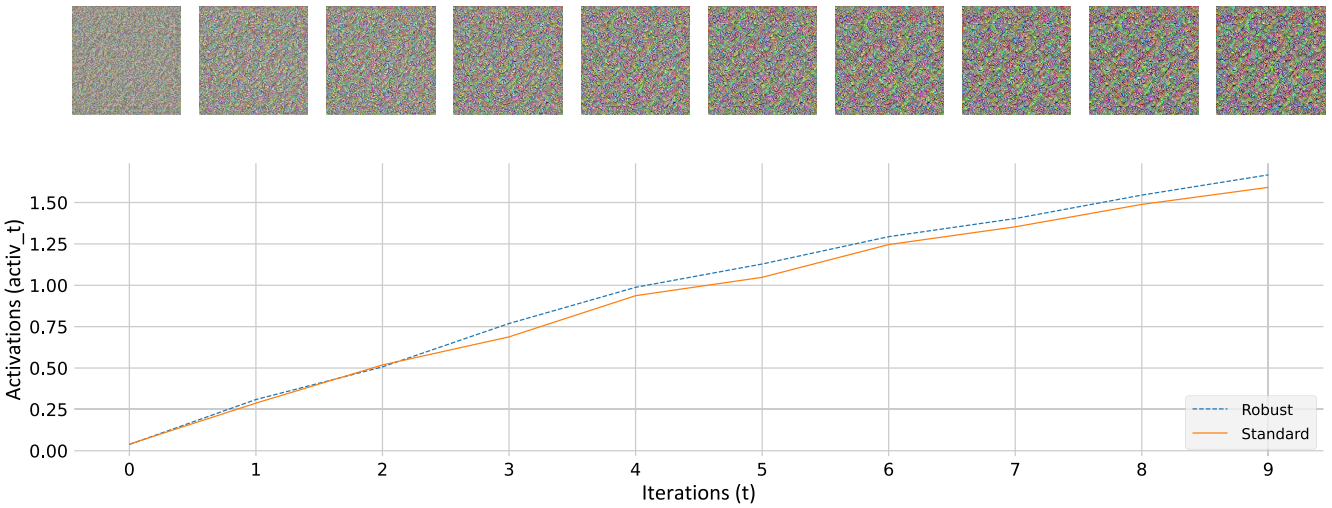


FIGURE 4. Example feature generation process showing how points on the activation plot relate to the FV images generated using Vanilla AM on ResNet. The FVs at the top correspond to x_t, where t is equal to each iteration in the graph below them.

over multiple iterations. The FVs are then evaluated for their activation scores. This process is reflected in Fig. 4 where the FVs (using Vanilla AM) and their corresponding activations are plotted together for a standard and robust model (ResNet). In Fig. 4, the activation curves are relative to the FVs that were generated at that specific iteration. For example, iteration 3 in Fig. 4 corresponds to an activation score of 0.75 for the robust model. This process is also depicted in Fig. 3, where iterations correspond to t and activations correspond to $activ_t$. The individual FVs for each iteration (x_t) are presented at the top of Fig. 4. These FVs are then individually input back into the model to get the activation score shown in the activation curve below each FV.

The activation curves generated by these FVs are useful to understand, as they are the basis of the IntActEval method. We present the activation results for the two best performing FV methods in Fig. 5 for GoogLeNet. The IntActEval score is calculated by taking the AUC of the activation plots, where the AUC of Fig. 5a and 5b correspond to the bar plots in Fig. 5c for Gaussian Noise and Vanilla respectively. For each activation curve, we use the trapezoidal rule to calculate AUC. The process is then repeated using the standard and robust models for each FV method. The FVs were generated 10 times for each method to get an average activation at each step. We also show the statistical significance of each method using the 95% confidence interval (CI) as seen in Fig. 5.

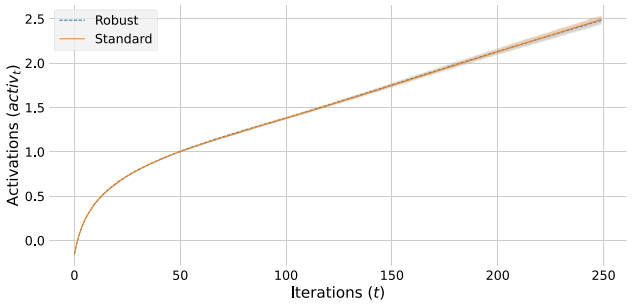
Examining the IntActEval results in Fig. 5c, we observe no statistically significant difference between the robust and standard trained models for each FV method. This finding holds true across all five FV methods for the ResNet model. However, when comparing the five FV methods themselves, we do find statistically significant differences. For example, there is a glaring difference between the FP and TR methods for both the standard and robust cases. These observations are consistent across all tested models, as we will further discuss in the following subsection.

B. INTACTEVAL (AUC) RESULTS

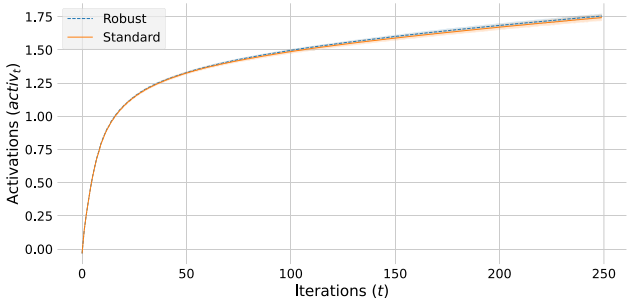
The IntActEval score, also referred to as AUC or faithfulness score, is derived from the FV activation scores, as described in the preceding subsection. For our tests, we first grouped the results by model type (standard, robust) and then by FV type (Gaussian, TRNFP, Vanilla, TR, and FP). Each FV was generated 10 times for each method and model, and the averages were calculated. This enabled us to compute confidence intervals and perform statistical analysis.

Using the IntActEval results presented in Fig. 6, we conduct a comparative analysis of different methods to identify the most suitable FV for enhancing model trust by determining which method is most faithful. The larger the IntActEval score is, the more faithful the FV method is determined to be. For the IntActEval results presented in these bar plots, we found no statistically significant difference between the robust and standard trained models for any particular FV method. This shows that adversarial training using the PGD attack does not affect the faithfulness of FVs. More testing is needed to confirm whether this is true for all robust training schema.

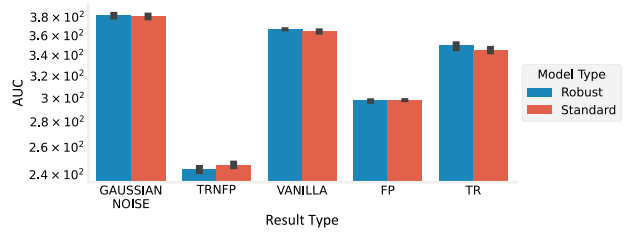
When comparing the IntActEval scores for the five FV methods, we did find statistically significant variation among them. This is clearly shown for all models in Fig. 6. Notably, the TR, FP, and TRNFP regularizers performed worse than the Vanilla and Gaussian Noise approaches. These lower-performing regularizers involved the reduction of noise in the FV image, suggesting that the model assigns higher activation scores to images with a high degree of noise. Furthermore, some of the stronger regularizers, such as TR, have been shown to produce more plausible results [17]. This could indicate that plausibility and faithfulness of FV



(a) Activations over iterations of Gaussian FV method for GoogLeNet.



(b) Activations over iterations of Vanilla FV method for GoogLeNet.



(c) AUC (IntActEval score) Comparison graph of each method for GoogLeNet

FIGURE 5. Result plots depict the activations versus AUC comparison for each FV method on GoogLeNet standard and robust. The bar plots for Gaussian Noise in (c) depicts the AUC of the activations in (a). Similarly, the AUC bar plots for Vanilla in (c) correspond to the activations in (b). All plots also show the 95% confidence interval (CI).

methods may be at odds with each other for these CNN models.

As illustrated in Fig. 6, a significant variance is observed in the relative performance of the TR and FP methods across different network architectures. For instance, the TR method yields a substantially higher AUC score for DenseNet, whereas the converse is true for SqueezeNet. We hypothesize that this discrepancy arises from an interaction between the filtering technique and the model-specific features learned by each network. If a method filters out features—such as specific textures, shapes, or patterns—that are integral to a particular model's predictive capability, a corresponding degradation in AUC will occur. Therefore, the observed fluctuation in performance is likely due to the unique feature dependencies of each model and which of those features are preserved by the TR and FP methods.

Next, we used the IntActEval score to rank the faithfulness of different methods. As shown in Fig. 7, the best performing FV methods were Gaussian Noise (regularized) and Vanilla (unregularized). These two methods performed similarly, with Gaussian Noise performing the best for five of the seven total models, and Vanilla performing best for the two remaining models. The number of iterations used to generate the explanation also affects the results, so this factor should be considered when choosing an FV method.

The third and fourth-ranked methods were FP and TR, respectively, with both methods performing nearly identically, except FP performed better on one additional model than TR. The worst-performing method in terms of faithfulness was TRNFP. For the models and datasets tested in this paper, we suggest that Gaussian or Vanilla FV methods should be used when faithful explanations are needed. If a more visually plausible but less faithful explanation is needed, other methods can be considered.

C. VISUAL COMPARISON OF FV METHODS

This subsection presents a visual comparison of FV methods by analyzing the explanations depicted in Figs. 8 through 14. All examples were generated after 250 ($n = 250$) iterations. Based on the IntActEval results, we can draw conclusions about how visual aspects of the FV methods relate to the faithfulness and the learned features of the model.

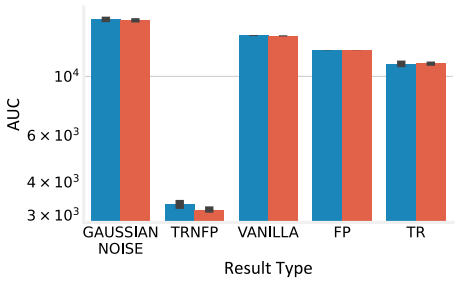
A key observation from the figures is that the Vanilla and Gaussian methods exhibit substantially more noise compared to TRNFP, FP, and TR. For example, Fig. 10 demonstrates that Vanilla and Gaussian produce irregular pixel-wise color transitions, whereas the other three methods display smoother and more consistent gradients. Notably, these latter methods also achieve higher faithfulness scores, as established earlier in this section.

We argue that the models in this study learn noise as a discriminative feature for decision-making. This aligns with prior work demonstrating that neural networks can exploit adversarial noise as a deceptive feature, leading to erroneous predictions [84].

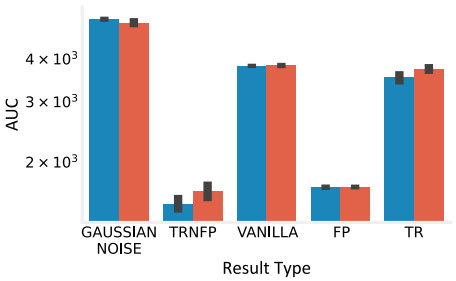
D. FV METHOD USAGE RECOMMENDATIONS

It is important that we use faithful explanations and robust models to achieve the most trustworthy and plausible results. It is critical that our explanations be true to the model while still depicting relevant features. Robustness plays a large role in plausibility and thus should be a significant consideration. It is important that we use faithful explanations and robust models for the most trustworthy and plausible results.

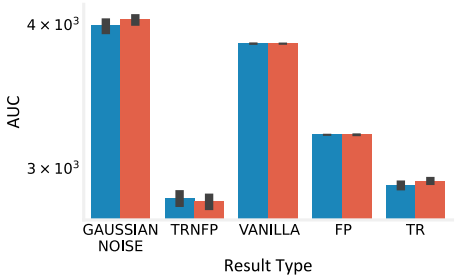
Robust models generally produce more plausible results [17], [51], [85], [86], indicating that they likely recognize more relevant features. For example, robust models generally perform better on out-of-distribution data, and perform far better than their non-robust counterparts when faced with high-frequency noise or adversarial attacks. Therefore, robust



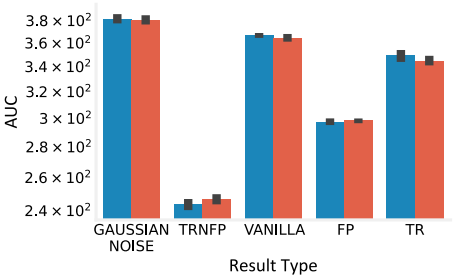
(a) AlexNet standard versus robust CIFAR10



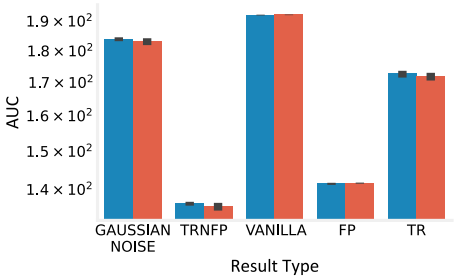
(b) DenseNet standard versus robust CIFAR10



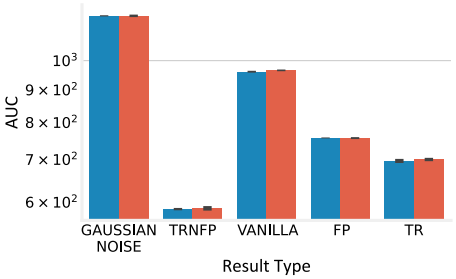
(c) EfficientNet standard versus robust CIFAR10



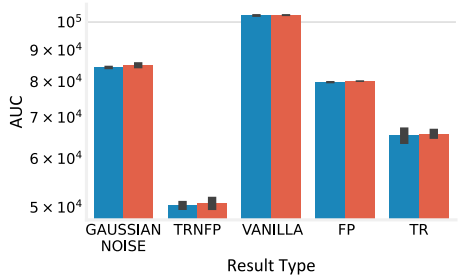
(d) GoogLeNet standard versus robust CIFAR10



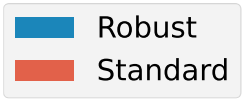
(e) MobileNetV2 standard versus robust CIFAR10



(f) ResNet standard versus robust CIFAR10



(g) SqueezeNet standard versus robust CIFAR10



(h) Legend for standard versus robust plots

FIGURE 6. Comparison of IntActEval scores across different models evaluated using various FV techniques, demonstrating the difference between robust and standard methods. The plots are in logarithmic scale to fit the large variance in activation scores. The error bars depict a 95% confidence interval (CI).

models should be used in most situations when possible, especially for explainability.

Faithful explanations are vital for understanding what the model is *really* looking at. While regularized FV methods have a tendency to produce more *plausible* explanations [38], our IntActEval results demonstrate that highly regularized techniques (e.g., TR, FP, and TRNFP) tend to sacrifice faithfulness. Conversely, robust models inherently yield more plausible visualizations without compromising faithfulness.

To balance plausibility and faithfulness, we recommend pairing the most faithful FV methods with robust models. In our experiments, Vanilla and Gaussian FVs exhibited comparable performance. However, the Gaussian approach resulted in more pronounced noise artifacts, thereby making

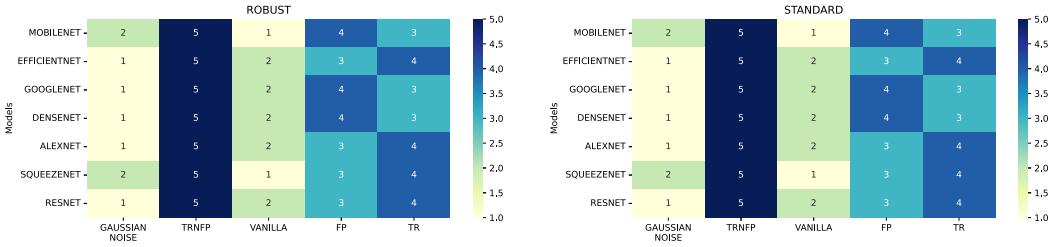
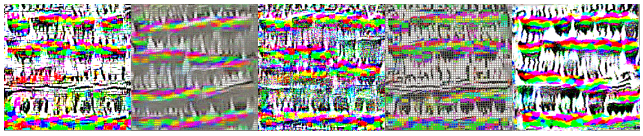
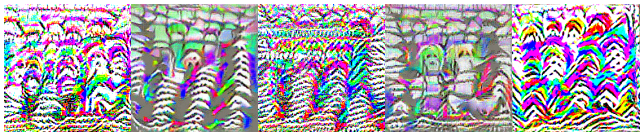


FIGURE 7. Rank comparison of IntActEval scores from Fig. 6 across each model. The methods are ranked from most (1) to least faithful (5).



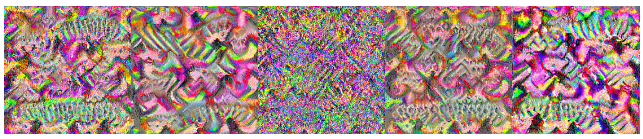
(a) Vanilla (b) TRNFP (c) Gaussian (d) FP (e) TR

FIGURE 8. FV methods on standardly trained AlexNet generated via AM from the final convolutional layer.



(a) Vanilla (b) TRNFP (c) Gaussian (d) FP (e) TR

FIGURE 9. FV methods on standardly trained DenseNet generated via AM from the final convolutional layer.

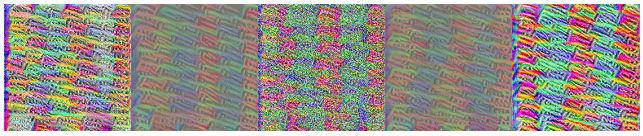


(a) Vanilla (b) TRNFP (c) Gaussian (d) FP (e) TR

FIGURE 10. FV methods on standardly trained efficientnet generated via AM from the final convolutional layer.

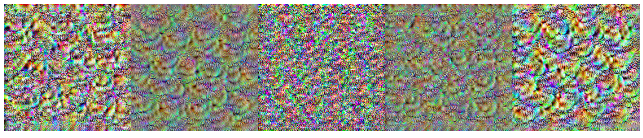
(a) Vanilla (b) TRNFP (c) Gaussian (d) FP (e) TR

FIGURE 11. FV methods on standardly trained GoogLeNet generated via AM from the final convolutional layer.



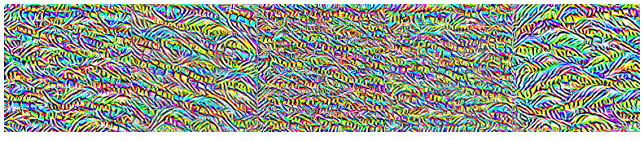
(a) Vanilla (b) TRNFP (c) Gaussian (d) FP (e) TR

FIGURE 12. FV methods on standardly trained MobileNetV2 generated via AM from the final convolutional layer.



(a) Vanilla (b) TRNFP (c) Gaussian (d) FP (e) TR

FIGURE 13. FV methods on standardly trained ResNet18 generated via AM from the final convolutional layer.



(a) Vanilla (b) TRNFP (c) Gaussian (d) FP (e) TR

FIGURE 14. FV methods on standardly trained squeezenet generated via AM from the final convolutional layer.

Vanilla the preferable choice. Furthermore, we advocate for the application of IntActEval to assess prospective FV methods, ensuring their reliability and faithfulness are consistent with the specific use cases.

E. PRODUCING FAITHFUL EXPLANATION

Our results found that robustness is not correlated to faithfulness with any statistical significance, based on the 95% confidence intervals (CI) shown in Fig. 6. Given our objective to maximize both plausibility and faithfulness, we propose that future research in this domain should consistently employ the use of robustly trained models or robust architectures to achieve the most plausible results without sacrificing faithfulness. These models should be used alongside either Gaussian or Vanilla methods of FV, which were found to produce the most faithful explanations.

F. FUTURE WORK

To advance the development of truly reliable and interpretable AI, future work must extend the IntActEval framework along three critical research paths. First, to establish IntActEval as a standard benchmark, its applicability must be broadened by evaluating more diverse model architectures, such as Vision Transformers (ViTs), and additional FV methods, including learned priors and dataset examples. This comprehensive

analysis is necessary to validate how faithfulness varies across the wider machine learning landscape. Second, a systematic investigation into the relationship between robustness and faithfulness is required. Applying IntActEval to models trained with various adversarial and data augmentation techniques is crucial to identify the conditions under which a correlation emerges, as understanding this interplay is fundamental to building models that are both secure and transparent. Finally, to be practically useful, explanations must be both faithful and intelligible. Therefore, the final, essential path is to quantify the observed trade-off between faithfulness and plausibility by integrating a quantitative plausibility metric, thereby creating a holistic evaluation that jointly measures a visualization's fidelity to the model and its human interpretability.

VI. SUMMARY AND CONCLUSION

The results of our novel evaluation revealed that there are notable differences in the faithfulness of different FV techniques. We found that Vanilla and Gaussian Noise FV methods produced the most faithful explanations overall, and this was consistent across all models tested. We also showed that robust training did not have a statistically significant effect on faithfulness. We noted earlier in the paper that robust models produce more plausible results on the tested data. Therefore, we recommend that the community use robust training to achieve the most plausible explanations without sacrificing faithfulness.

We also found that the most faithful methods also contained the most noise. This likely indicates that models which use a CNN architecture are learning to identify the presence of noise as an important feature when making a decision.

Trust is necessary when generating explanations, which is why the faithfulness of FV methods *must* be considered when discussing and developing these techniques in the future. Our novel faithfulness metric of IntActEval allows for this to be quantitatively assessed. Having an explanation be faithful is paramount if an approach is to be used for understanding deep learning models and what they truly learn.

ACKNOWLEDGMENT

(Bradley Gathers, Ian E. Nielsen, and Keith W. Soules contributed equally to this work.)

REFERENCES

[1] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: A survey," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8647–8695, Aug. 2023.

[2] F. Rundo, F. Trenta, A. L. di Stallo, and S. Battiato, "Machine learning for quantitative finance applications: A survey," *Appl. Sci.*, vol. 9, no. 24, p. 5574, Dec. 2019.

[3] T. M. Rausch, N. D. Derra, and L. Wolf, "Predicting online shopping cart abandonment with machine learning approaches," *Int. J. Market Res.*, vol. 64, no. 1, pp. 89–112, Jan. 2022.

[4] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 712–733, Feb. 2021.

[5] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," *Diagnostics*, vol. 12, no. 2, p. 237, Jan. 2022.

[6] D. Saraswat, P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, P. N. Bokoro, and R. Sharma, "Explainable AI for healthcare 5.0: Opportunities and challenges," *IEEE Access*, vol. 10, pp. 84486–84517, 2022.

[7] J. Chu, K. H. B. Leung, P. Snobelen, G. Nevils, I. R. Drennan, S. Cheskes, and T. C. Y. Chan, "Machine learning-based dispatch of drone-delivered defibrillators for out-of-hospital cardiac arrest," *Resuscitation*, vol. 162, pp. 120–127, May 2021.

[8] T. K. Balaji, C. S. R. Annavarapu, and A. Bablani, "Machine learning algorithms for social media analysis: A survey," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100395.

[9] F. Hussain, R. Hussain, S. A. Hassan, and E. Hossain, "Machine learning in IoT security: Current solutions and future challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1686–1721, 3rd Quart., 2020.

[10] J. Praful Bharadiya, "A comprehensive survey of deep learning techniques natural language processing," *Eur. J. Technol.*, vol. 7, no. 1, pp. 58–66, May 2023.

[11] A. A. Khan, A. A. Laghari, and S. A. Awan, "Machine learning in computer vision: A review," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 8, no. 32, 2018, Art. no. 169418.

[12] C. Maathuis, "On explainable AI solutions for targeting in cyber military operations," in *Proc. Int. Conf. Cyber Warfare Secur.*, 2022, vol. 17, no. 1, pp. 166–175.

[13] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," 2021, *arXiv:2112.11561*.

[14] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022.

[15] L. Nannini, J. M. Alonso-Moral, A. Catalá, M. Lama, and S. Barro, "Operationalizing explainable artificial intelligence in the European union regulatory ecosystem," *IEEE Intell. Syst.*, vol. 39, no. 4, pp. 37–48, Jul. 2024.

[16] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision making and a 'right to explanation,'" *AI Mag.*, vol. 38, no. 3, pp. 50–57, Sep. 2017.

[17] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, no. 11, 2017. [Online]. Available: https://distill.pub/2017/feature-visualization

[18] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Gradient-based attribution methods," in *Explainable AI: Interpreting, Explaining Visualizing Deep Learn.*, 2019, pp. 169–191.

[19] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.

[20] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[21] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 9737–9748. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf

[22] I. E. Nielsen, R. P. Ramachandran, N. Bouaynaya, H. M. Fathallah-Shaykh, and G. Rasool, "EvalAttAI: A holistic approach to evaluating attribution maps in robust and non-robust models," *IEEE Access*, vol. 11, pp. 82556–82569, 2023.

[23] N. Hama, M. Mase, and A. B. Owen, "Deletion and insertion tests in regression models," 2022, *arXiv:2205.12423*.

[24] M. N. Vu, T. D. Nguyen, N. Phan, R. Gera, and M. T. Thai, "C-eval: A unified metric to evaluate feature-based explanations via perturbation," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 927–937.

[25] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Univ. Montreal*, vol. 1341, no. 3, p. 1, May 2009.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5188–5196.

[28] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," 2015, *arXiv:1506.06579*.

[29] A. Mordvintsev, C. Olah, and M. Tyka, "Inceptionism: Going deeper into neural networks," Google Res. Blog, Tech. Rep., 2015. [Online]. Available: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

[30] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 427–436.

[31] S. Krishna, T. Han, A. Gu, S. Wu, S. Jabbari, and H. Lakkaraju, "The disagreement problem in explainable machine learning: A practitioner's perspective," *Trans. Mach. Learn. Res.*, 2024. [Online]. Available: https://openreview.net/forum?id=jESY2WTZCe

[32] T. Han, S. Srinivas, and H. Lakkaraju, "Which explanation should i choose? A function approximation perspective to characterizing post hoc explanations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 5256–5268. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/22b111819c74453837899689166c4cf9-Paper-Conference.pdf

[33] D. Dera, N. C. Bouaynaya, G. Rasool, R. Shterenberg, and H. M. Fathallah-Shaykh, "PremiUm-CNN: Propagating uncertainty towards robust convolutional neural networks," *IEEE Trans. Signal Process.*, vol. 69, pp. 4669–4684, 2021.

[34] D. Dera, G. Rasool, and N. Bouaynaya, "Extended variational inference for propagating uncertainty in convolutional neural networks," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2019, pp. 1–6.

[35] G. Carannante, N. C. Bouaynaya, D. Dera, H. M. Fathallah-Shaykh, and G. Rasool, "SUPER-Net: Trustworthy image segmentation via uncertainty propagation in encoder–decoder networks," 2021, *arXiv:2111.05978*.

[36] I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, and N. C. Bouaynaya, "Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks," *IEEE Signal Process. Mag.*, vol. 39, no. 4, pp. 73–84, Jul. 2022.

[37] M. Nourelahi, L. Kotthoff, P. Chen, and A. Nguyen, "How explainable are adversarially-robust CNNs?" 2022, *arXiv:2205.13042*.

[38] A. Nguyen, J. Yosinski, and J. Clune, "Understanding neural networks via feature visualization: A survey," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 2019, pp. 55–76.

[39] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah, "Activation atlas," *Distill*, vol. 4, no. 3, p. 15, Mar. 2019.

[40] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. 2nd Int. Conf. Learn. Represent.*, 2022.

[41] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: Conditional iterative generation of images in latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3510–3520.

[42] A. S. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2022, pp. 3387–3395.

[43] A. Ghiasi, H. Kazemi, E. Borgnia, S. Reich, M. Shu, M. Goldblum, A. G. Wilson, and T. Goldstein, "What do vision transformers learn? A visual exploration," 2022, *arXiv:2212.06727*.

[44] G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, and C. Olah, "Multimodal neurons in artificial neural networks," *Distill*, vol. 6, no. 3, p. 30, Mar. 2021.

[45] M. Tyka. (2016). *Class Visualization With Bilateral Filters*. [Online]. Available: https://mtyka.github.io/deepdream/2016/02/05/bilateral-class-vis.html

[46] B. Heinzerling and M. Strube, "Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation," Tech. Rep., 2019.

[47] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[48] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, 2022.

[49] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1103, Oct. 2018.

[50] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[51] I. E. Nielsen, E. Grundeland, J. Snedeker, R. P. Ramachandran, and G. Rasool, "Targeted background removal creates interpretable feature visualizations," in *Proc. IEEE 66th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2023, pp. 1050–1054.

[52] P. Chalasani, J. Chen, A. R. Chowdhury, X. Wu, and S. Jha, "Concise explanations of neural networks using adversarial training," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, Jun. 2020, pp. 1383–1391. [Online]. Available: https://proceedings.mlr.press/v119/chalasani20a.html

[53] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3319–3327.

[54] G. V. Nguyen, D. Kim, and A. Nguyen, "The effectiveness of feature attribution methods and its correlation with automatic evaluation scores," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2022, pp. 26422–26436. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/de043a5e421240eb846da8effe472ff1-Paper.pdf

[55] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*. Cham, Switzerland: Springer, 2014, pp. 818–833. Accessed: D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars.

[56] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[57] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, Aug. 2022, pp. 3319–3328. [Online]. Available: https://proceedings.mlr.press/v70/sundararajan17a.html

[58] Y. Zhang, Y. Li, H. Brown, M. Rezaei, B. Bischl, P. Torr, A. Khakzar, and K. Kawaguchi, "Attributionlab: Faithfulness of feature attribution under controllable environments," Tech. Rep., 2024. [Online]. Available: https://openreview.net/forum?id=zhINOCrrqI

[59] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2022, pp. 9505–9515. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf

[60] S. Palacio, F. Raue, T. Karayil, J. Hees, and A. Dengel, "Iteroar: Quantifying the interpretation of feature importance methods," in *Proc. 35th Pre-Registration Workshop (NeurIPS)*, 2021.

[61] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2131–2145, Sep. 2019.

[62] A. M. Salih, I. B. Galazzo, P. Gkontra, E. Rauseo, A. M. Lee, K. Lekadir, P. Radeva, S. E. Petersen, and G. Menegaz, "A review of evaluation approaches for explainable AI with applications in cardiology," *Artif. Intell. Rev.*, vol. 57, no. 9, p. 240, Aug. 2024.

[63] J. Kim, H. Maathuis, and D. Sent, "Human-centered evaluation of explainable AI applications: A systematic review," *Frontiers Artif. Intell.*, vol. 7, Oct. 2024.

[64] X. Lu and J. Ma, "Does faithfulness conflict with plausibility? An empirical study in explainable AI across NLP tasks," 2024, *arXiv:2404.00140*.

[65] S. Rao, M. Böhle, and B. Schiele, "Towards better understanding attribution methods," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10213–10222.

[66] M. Pawlicki, A. Pawlicka, F. Uccello, S. Szelest, S. D'Antonio, R. Kozik, and M. Choraś, "Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination," *Neurocomputing*, vol. 602, Oct. 2024, Art. no. 128282.

[67] A. Nguyen, J. Yosinski, and J. Clune, "Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks," 2016, *arXiv:1602.03616*.

[68] C. Cortes, M. Mohri, and A. Rostamizadeh, "L 2 regularization for learning kernels," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 109–116.

[69] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[70] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 6105–6114.
[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 25, 2012.
[72] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," 2014, *arXiv:1404.5997*.
[73] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[74] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[75] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*.
[76] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[77] S. Marcel and Y. Rodriguez, "Torchvision the machine-vision package of torch," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 1485–1488.
[78] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Computer science, Univ. Toronto, Toronto, ON, Canada, 2009.
[79] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019, *arXiv:1706.06083*.
[80] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.
[81] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.
[82] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.
[83] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," 2019, *arXiv:1911.02685*.
[84] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2022, pp. 125–136. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf
[85] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, and A. Madry, "Adversarial robustness as a prior for learned representations," 2019, *arXiv:1906.00945*.
[86] S. Sietzen, M. Lechner, J. Borowski, R. Hasani, and M. Waldner, "Interactive analysis of CNN robustness," in *Computer Graphics Forum*. Hoboken, NJ, USA: Wiley, vol. vol. 40, 2021, pp. 253–264.

BRADLEY GATHERS received the B.S. degree in computer science in 2024. He is currently pursuing the M.S. degree in computer science with Rowan University.

From 2022 to 2024, he was with the Rotary and Missions Systems Department, Lockheed Martin, as a Software Engineer. He has been a Software Engineer with JPMorgan Chase and Co., since 2024. He was also the Judge in 2024 as part of the locally run IEEE Hackathon (ProfHacks). His research interests and efforts are in eXplainable artificial intelligence (XAI), focusing on computer vision tasks, with a particular focus on the topics of robustness, plausibility, and faithfulness. Since 2022, he has been assisting Junior and Senior undergraduate student researchers alongside Dr. Ian Nielsen, Keith Soules, and Dr. Ravi Ramachandran via the Rowan University Engineering Clinic programs. Additionally, since 2022, he has been assisting Dr. Ian Nielsen, Keith Soules, and Dr. Ravi Ramachandran in the implementation, design, and theorization of their ongoing research efforts.

IAN E. NIELSEN (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Rowan University, Glassboro, NJ, USA, in 2025. He is currently a Founding Researcher with 2nd Set AI, where he develops and benchmarks generative models for image and video. As a Ph.D. student, he conducted research at Rowan's Machine and Artificial Intelligence VR Center (MAVRC), and co-mentored undergraduate students in machine learning through the university's engineering clinic program. His work has been published in journals and conferences, including *IEEE Signal Processing Magazine*, *Circuits, Systems, and Signal Processing* (CSSP), the IEEE International Midwest Symposium on Circuits and Systems (MWSCAS), and IEEE Access. His research interests include robust machine learning, explainable artificial intelligence (XAI), computer vision, and object detection. During his graduate studies, he was a recipient of U.S. Department of Education GAANN Teaching Fellowship.

KEITH W. SOULES received the B.S. and M.S. degrees from Rowan University, in 2023 and 2024, respectively, where he is currently pursuing the Ph.D. degree. He is a Teaching Assistant and a Research Assistant with Rowan University. He conducts research on eXplainable artificial intelligence (XAI), multimodal object detection, and predictive maintenance of wind farms through object detection techniques. He also teaches and conducts XAI research with a team of undergraduate students through the Rowan University Engineering Clinic Program. He was a recipient of the NJ Wind Institute Fellowship.

OZAN TEKBEN, photograph and biography not available at the time of publication.

RAVI P. RAMACHANDRAN (Senior Member, IEEE) received the B.Eng. degree (Hons.) from Concordia University, in 1984, the M.Eng. degree from McGill University, in 1986, and the Ph.D. degree from McGill University, in 1990. From October 1990 to December 1992, he was with the Speech Research Department, AT&T Bell Laboratories. From January 1993 to August 1997, he was a Research Assistant Professor with Rutgers University. He was also a Senior Speech Scientist with T-Netix, from July 1996 to August 1997. Since September 1997, he has been with the Department of Electrical and Computer Engineering, Rowan University, where he has been a Professor, since September 2006. He served as a consultant to T-Netix, Avenir Inc., Motorola, and FocalCool. His research interests include digital signal processing, speech processing, biometrics, pattern recognition, machine learning, and filter design. From September 2002 to September 2005, he was an Associate Editor of IEEE Transactions on Speech and Audio Processing and was on the Speech Technical Committee of the IEEE Signal Processing Society. From September 2000 to December 2015, he was on the Editorial Board of the *IEEE Circuits and Systems Magazine*. Since May 2002, he has been on the Digital Signal Processing Technical Committee of the IEEE Circuits and Systems Society. Since May 2012, he has been on the Education and Outreach Technical Committee for the IEEE Circuits and Systems Society. He is also an Associate Editor of *Circuits, Systems, and Signal Processing*.

NIDHAL CARLA BOUAYNAYA (Member, IEEE) received the M.S. degree in pure mathematics and the Ph.D. degree in electrical and computer engineering (ECE) from the University of Illinois at Chicago.

She is currently a Professor of ECE and the Director of Rowan's Artificial Intelligence Laboratory (RAIL). She is also the Associate Dean of the Research and Graduate Studies at the Henry M. Rowan College of Engineering. Previously, she was a Faculty Member with the University of Arkansas at Little Rock. She has co-authored more than 100 referred journal articles, book chapters, and conference proceedings, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE SIGNAL PROCESSING LETTERS, *IEEE Signal Processing Magazine*, and *PLOS Medicine*. Her research is primarily funded by the National Science Foundation (NSF CCF, NSF ACI, NSF DUE, NSF I-Corps, NSF ECCS, NSF OAC, and NSF HRD), The National Institutes of Health (NIH), U.S. Department of Education (USED), New Jersey Department of Transportation (NJ DoT), U.S. Department of Agriculture (USDA), the Federal Aviation Administration (FAA), Lockheed Martin Inc., and other industry. She is also interested in entrepreneurial endeavors. In 2017, she co-founded and is the Chief Executive Officer (CEO) of MRIMATH, LLC, a start-up company that uses artificial intelligence to improve patient oncology outcomes and treatment response. MRIMath is funded by the NIH SBIR Program. Her research interests include big data analytics, machine learning, artificial intelligence, and mathematical optimization. She won numerous best paper awards, the most recent being at the 2019 IEEE International Workshop on Machine Learning for Signal Processing. She is also the winner of the Top Algorithm at the 2016 Multinomial Brain Tumor Segmentation Challenge (BRATS). She has been honored with numerous research and teaching awards, including the Rowan Research Achievement Award in 2017 and The University of Arkansas at Little Rock Faculty Excellence Award in Research.

HASSAN M. FATHALLAH-SHAYKH (Member, IEEE) received the M.D. degree from American University of Beirut, and the Ph.D. degree in pure mathematics from the University of Illinois at Chicago. He is currently a Professor with The University of Alabama at Birmingham, with appointments in neurology and mathematics.

GHULAM RASOOL (Member, IEEE) received the B.S. degree in mechanical engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2000, the M.S. degree in computer engineering from the Center for Advanced Studies in Engineering (CASE), Pakistan, in 2010, and the Ph.D. degree in systems engineering from the University of Arkansas at Little Rock, in 2014. He was a Postdoctoral Fellow with the Rehabilitation Institute of Chicago and Northwestern University, from 2014 to 2016. Before joining Moffitt, he was an Assistant Professor with the Department of Electrical and Computer Engineering, Rowan University. He is currently an Assistant Member with the Department of Machine Learning, H. Lee Moffitt Cancer Center, and the Research Institute, Tampa, FL, USA. His current research interests include building trustworthy multimodal machine learning and artificial intelligence models for cancer diagnosis, treatment planning, and risk assessment. His research efforts are funded by the National Science Foundation (NSF), the National Institutes of Health (NIH), and the Moffitt Cancer Center.