Quantitative Evaluation of Video Explainability Methods via Anomaly Localization

Xinyue Zhang[1] Boris Joukovsky[1,2] Nikos Deligiannis[1,2]

[1]Department of Electronics and Informatics, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
[2]imec, Kapeldreef 75, B-3001 Leuven, Belgium

***Abstract*—This paper presents a study of explainable AI methods applied to video anomaly detection. Specifically, we put forward a multidimensional evaluation protocol to evaluate attribution methods by considering the correctness of the explanations, their plausibility with respect to ground-truth anomaly data, and the robustness of explanations across multiple time frames. We evaluate these metrics on common gradient-based and perturbation-based explanation techniques, which we use to explain a 3D convolutional neural network trained on real video data. Our results show that using specific methods generally leads to trade-offs between the explanation performance and the higher computational cost related to video data. In particular, gradient-based methods achieve higher robustness across multiple frames, whereas perturbation methods achieve higher model fidelity scores.**

***Index Terms*—deep learning, video anomaly detection, explainability methods, quantitative evaluation**

I. Introduction

Anomaly detection in video is one of the topics of interest in video recognition using deep neural networks [1]. This task is challenging due to the wide variety of possible behaviors that can be considered abnormal, and these actions often happen scarcely and with a short duration. 3D Convolutional Neural Networks (3DCNNs) constitute one of the main category of models that can be applied to video: they extract features both in the spatial and temporal dimensions using 3D convolution kernels [2]–[5] and show great performance on the anomaly detection task [6].

Despite the success of deep models, they also behave like black-boxes, since their complex feature extraction process makes it difficult to understand how they operate internally. Explainable AI (XAI) techniques aim at addressing this limitation by designing methods to understand which (parts of) inputs most likely cause a given prediction; we refer to [7] for a review on different methods, and in addition, visual and temporal cues in the input to video model are often hard to disentangle [8]. More generally, the use of XAI methods raises the concerns of whether they can correctly and faithfully reflect the model behaviour and whether the explanations are consistent for every frame in the video. Therefore, in this paper, we use the practical case of video anomaly detection to study various explanation methods applied to a trained 3D convolutional network [5]. Using quantitative experiments, we seek to study (1) whether the explanation methods faithfully reflect the decision of model, (2) if the explanations are plausible and understandable for users, and (3) whether the methods perform stably in time, with a consistent confidence for all video frames. Our methodology consists of the following:

- We implement and train the existing I3D model [5] on the USCD video anomaly detection dataset [9] to predict anomalies.
- We use gradient-based and perturbation-based explainability techniques to visualize and localize the anomalies via heatmaps, thereby providing a first qualitative assessment of the methods.
- We systematically assess the explainability techniques via quantitative metrics. Specifically, we employ existing metrics to measure the faithfulness of the explanation w.r.t the model (referred to as *model-centric*), we evaluate their plausibility via unsupervised localization of the anomalies (referred to as *human-centric*), we introduce a measure robustness for the explanations across time, and compare the computational costs of the different methods.

The remainder of the paper is organized as follows: Section II presents the background on 3D convolutional neural networks and the different explainability methods considered in this work. We describe the proposed evaluation metrics in Section III. Our experimental results and findings are presented in Section IV, and Section V concludes the paper.

II. Related work

A. I3D Model and Anomaly Detection

3DCNNs [2] use cube-shaped convolutional kernels on to capture spatio-temporal features. The model considered in this work is the Two-Stream Inflated 3D ConvNet (I3D) model [5], which is built by inflating a 2D Inception v1 network [10] pre-trained on the ImageNet dataset. Inception v1 uses inception modules to alleviate the problems of explosion or vanishing of gradients, heavy computation load and overfitting in a large-scale network.

In surveillance videos, anomalies usually refer to the unnatural activities such as theft, arson nearby buildings, intrusions, etc. In the domain of deep learning, it is often viewed as a video classification task [11]. When anomalies are not differentiated, the problem boils down to a binary classification or detection task, with successful approaches that include 3DCNNs trained both in supervised and semi-supervised settings [1], [12]. In the context of anomaly detection, models

should be able to predict on temporal and spatial dimensions respectively, that means models need to decide anomalies on frame level (i.e., temporal localization), and distinguish which specific pixels on the frames show the abnormal behaviors (i.e., spatial localization on pixel-level) [13].

B. Explainability Methods

In this work, we consider two categories of explainability techniques, which are the gradient-based and perturbation-based methods. Other methods involving deep networks as explanation models (e.g.: transformers-based explainer [14]) are not considered here.

Gradient-based methods: These techniques exploit the local gradients of the model function to identify the most important inputs. The Saliency Maps method [15] uses gradients obtained via the backpropagation pass to highlight the most salient features. Similarly, Guided BackPropagation [16] pools the positive gradients at every layer during the backpropagation pass, which improves the quality of the saliency map. However, these approaches are sensitive to gradient vanishing and saturation regimes, which can lead to incomplete explanations. Integrated Gradients [17] alleviates these issues by computing an approximation of the straight-path integral of gradients, requiring more forward passes as the sampling times increases. SmoothGrad [18] smooths the noisy gradients effectively by averaging over multiple passes.

Perturbation-based methods: These methods are essentially model-agnostic and rely on changes of the prediction w.r.t. input perturbations. However, they usually require carefully designed sampling schemes and more forward model evaluations to accurately estimate the input importance scores. The Occlusion method [19] considers individual pixels or patches as perturbation units, and the associated attributions are directly given by the variation in output score as they are removed. Nevertheless, this approach has a high computation cost due to its exhaustive search nature. LIME [20] strongly reduces the amount of model evaluations by removing multiple patches at a time: the attributions are then obtained by training and interpretable local surrogate model (e.g., LASSO) on the observed variations. Additionally during regression, samples are weighted according to their similarity with the original input. KernelSHAP [21] is essentially a generalization of LIME, where the similarity kernel used to weight the training samples is replaced by Shapely values-based coefficients, which are used to redistribute feature contributions according to game theory-based considerations.

To evaluate explainable AI methods is a necessary step to ensure that they are trustworthy [22]. However, it a multi-dimensional problem that cannot be solved using a single metric or user study. For instance the co-12 properties [23] were proposed to comprehensively evaluate explainability methods on different aspects. For instance, from a content perspective, it is important to consider whether explanations can correctly, completely reflect the model predictions. From a presentation perspective, the explanation should be compact, avoiding large or sparse presentations. From a user perspective, the explanation should be plausible or interactable.

III. Proposed Evaluation Protocol

In this work, we consider a series of quantitative metrics for different aspects of video explanations in anomaly detection. First, we propose a group of *human-centric* metrics to compare an explanation map with ground-truth labels, that is, by viewing the explanation process as an unsupervised localization method to isolate the anomalies, which are then compared with human-annotated binary masks. This scheme can be used to compare various explainability methods together, but it cannot fully reflect the quality of an explanation since it does not account for the faithfulness of the explanation with respect to the model. Therefore, we complement this study with *model-centric* metrics, which consider the sensitivity of the model when parts of the data are removed according to the attributions. Additionally, we study the stability of the explanations in the temporal dimension, as well as their runtime complexities.

Human-centric metrics: We adopt the Intersection over Union (IoU), the F1 score and the area under curve (AUC) to measure the agreement between the explanations and ground-truth annotations of the anomalies. The metrics are computed respectively on clip-level (over entire single clip) and frame-level (on the frame that has maximum of attribution when anomalies entirely appear). Let M denote the ground-truth mask, E denote the explanation map and E_τ the map binarized according to the threshold τ. Then compute the IoU and F1 scores over the thresholds and select the maximum scores:

$$\mathrm{IoU} = \max_\tau \frac{|M \cap E_\tau|}{|M \cup E_\tau|}, \tag{1}$$

where the threshold τ iterates on the linear space ranging from 0 to the maximum value of attributions, N here represents the number of steps and is set to 50 in our experiments:

$$\tau_i = \frac{i}{N} \times E_{max}, \quad i \in \{0, 1, 2, ..., N\}. \tag{2}$$

The F1 score is usually defined as the harmonic mean of the precision and recall and is computed as:

$$\mathrm{F1} = \max_\tau \frac{2\mathrm{TP}_\tau}{(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})_\tau}, \tag{3}$$

where TP, FP, FN respectively stands for true positives, false positives and false negatives. High IoU and F1 scores indicate a good accuracy of anomalies localization. Different from the binarization step required for the computation of the IoU and F1 scores, the AUC accounts for the continuous nature of heatmaps; it is obtained by plotting the true positive rate (TPR) as a function of the false positive rate (FPR), and calculating the area under the resulting curve. Hence, the AUC provides a quality metric independent of the decision threshold.

Model-centric metrics: The model-centric metrics emphasize on output change when retaining or ablating important features and indicate how faithfully the explainability technique reflects the model decision. In this work, we adopt

the deletion and insertion metrics [24]. The deletion metric measures the drop in target probability when important pixels are gradually replaced by baseline values, while the insertion metric measures the rise of target probability when important pixels being added from the baseline. In our experiments, instead of directly computing the output probability, we calculate the output difference w.r.t the initial predicted probability. First, a softmax layer added on top of I3D model and the probability of anomaly is then explained. We normalize the resulting explanation maps after filtering out all negative attributions. The thresholds are generated according to Eq. (2) and are reversely iterated to compute the deletion and insertion metrics by averaging the measured differences.

Robustness metric: To assess the robustness of explainability techniques on the temporal dimension, we select the frames on which anomalies exist and compute a series of statistics over the frame-wise anomaly localization AUC based on the resulting attributions and ground-truth anomaly masks:

1) Mean value of AUC (mAUC): reflects the general anomaly-localization performance over a time period.
2) Standard deviation of AUC (stdAUC): indicates the stability of anomaly-localization along temporal dimension in a video.
3) Coefficient of variation (C_v): a combination of the above statistics:

$$C_v = \frac{\text{stdAUC}}{\text{mAUC}} \tag{4}$$

The coefficient of variation is traditionally used to measure the variability with respect to the population mean [25]. Thus, a small deviation combined with a large mean value implies that the explainability method has a stronger robustness along the temporal dimension.

IV. Experiments and Evaluation

In what follows, we report the training results of the I3D model on the UCSD video anomaly dataset [9], we demonstrate the visualizations of localizations of anomalies under different explainability methods, and we present quantitative evaluations on the anomaly localizations.

A. Training Results

In our experiments, we employ the I3D model with a single RGB stream, which is pre-trained on ImageNet before kernel inflation [5]. The model is trained on the two available scenes from the UCSD dataset (denoted by Ped1 and Ped2), from which we extract clips of 20 frames with a processed spatial size of 158×237 pixels. Each clip is labeled as anomalous if it contains at least one anomaly frame. The model is trained during 100 epochs with a batch size of 5 and using the Adam optimizer and the binary cross-entropy loss, with a fine-tuned learning rate of 0.001 and weight decay as 0.0001, which is decreased by a factor of 0.1 every 30 epochs. We apply data normalization and augmentation, including random cropping and rescaling. Our trained model achieves a test clip-wise accuracy of 88.86%, with an anomaly detection AUC of ROC of 0.9331 on the clip-level.

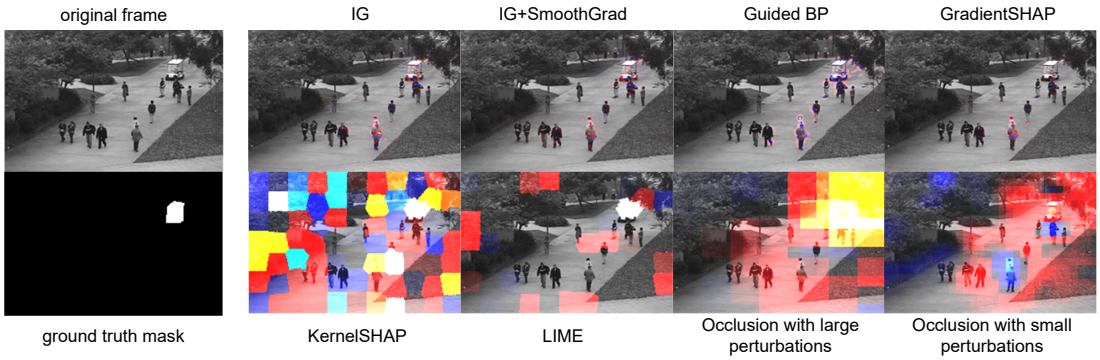
B. Explanations and Localization of Anomalies

We compare and analyse Integrated Gradients, SmoothGrad, Guided Backpropagation as for the gradient-based methods, and KernelSHAP, GradientSHAP, LIME, and Occlusion with small and with large perturbations as for the perturbation-based methods. We rely on the Captum library for their implementations [26].

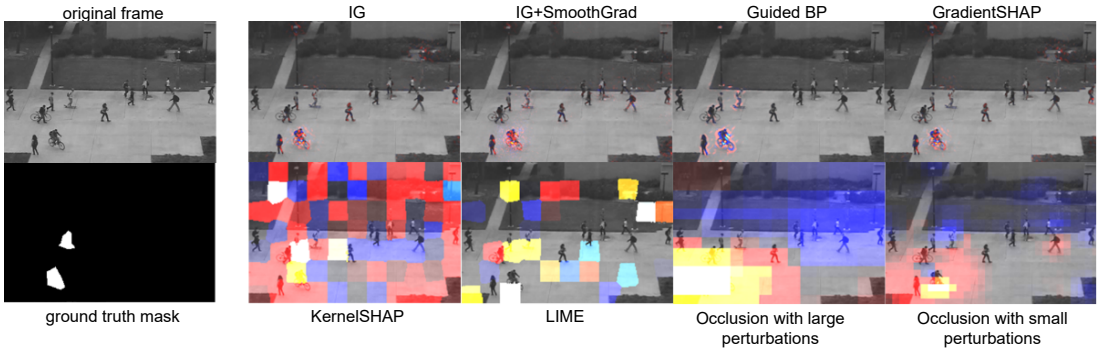
For Integrated Gradients, we set the number of integration steps to 30. The SmoothGrad method is also applied to Integrated Gradients (IG) with 30 steps and by sampling 5 noisy inputs according to a Gaussian distribution with standard deviation of 1.0. As for the Occlusion method, we employ a sliding window in two different settings: first, we use a large window of 10 frames and a spatial size 52×78 pixels with a stride of 10 frames and 32×32 pixels, which results in coarse-grain explanations, but with lower computation cost. Second, we use small perturbations with a sliding window of 4 frames and 16×24 pixels with a stride of 2 frames and 8×8 pixels. For LIME, we firstly create image 3D super-pixels with the SLIC algorithm, with 80 clusters and a compactness of 60. The surrogate model is a LASSO regressor, as commonly used with LIME, and we use a Gaussian similarity kernel with a width of 1,000 for the similarity function. As for the SHAP-related methods, KernelSHAP uses LIME to compute Shapley value more efficiently, while GradientSHAP uses expectations of gradients to approximate Shapley values [21]. Therefore, KernelSHAP adopts the same feature masks obtained in LIME, while GradientSHAP does not need any parameter settings. For the methods that require a baseline to fill the perturbed samples, we experiment with both Gaussian-blurred images with a kernel of size 9 and with black images. Nevertheless, we observed that the blurred baseline provide better explanation maps, hence we do not include the black baseline in this paper.

C. Qualitative Results

Figure 1 shows examples of heatmaps from two clips from both the Ped1 and Ped2 scenes. For clarity, we only display a single frame of the 20-frames long clips, along with the ground-truth mask of the anomalies. The attribution maps are superimposed to the frames, with positive and negative attributions rendered using red and blue color maps respectively. The anomalies are defined by the ground-truth masks provided in the UCSD dataset. In the first clip, the only anomaly is the white car in the upper right corner, whereas the second one features two anomalies, which are the biker and the skater located on the left side. However, in this same clip, the person pushing his bike is a misleading cue (based on the ground-truth mask). Gradient-based methods only highlight individual pixels, in contrast to perturbation-based methods where the attributions appear in patches, and the areas with the highest contributions can be recognized more easily with the presence of yellow patches. For gradient-based methods, we observed that IG and IG+SmoothGrad accurately localizes the anomalies, with less leakage on other objects than Guided Backpropagation (e.g.: irrelevant pedestrians are highlighted in first clip) since this last technique has a higher tendency to



(a) Ped1 scene.



(b) Ped2 scene.

Fig. 1: Visualization of attribution maps (Ped1 and Ped2 scenes). Most left column: original frame and ground truth mask. Right part: heatmaps for each explainability method.

behave like an edge detector [22]. For the perturbation-based methods, LIME concentrates the attributions more towards the ground-truth features compared to KernelSHAP. Occlusion with large perturbations leads to more coarse-grained features on the map, nevertheless, it has superior ability to cover anomalies than Occlusion with small perturbations.

D. Quantitative Results and Discussion

We present a quantitative evaluation based on the evaluation metrics described in Section III of the explanation methods. In order to avoid variability, all reported metrics are averaged over 40 clips selected randomly from the test set.

Human-centric metrics: We compute the IoU, F1 and AUC statistics w.r.t to ground-truth annotations for the anomalies on whole clips (clip level), and on the frame in each clip with maximum attribution and with an anomaly (max-attribution-frame level), since the localization performance can vary greatly between different frames and depending on the explanation method. Table I reports the results: as a general observation, the gradient methods tend to perform better than perturbations in either clip or frame-level case, with the exception of LIME that outperforms the other methods in the clip-level setting. In the max-attribution-frame setting, IG and IG+SmoothGrad perform the best.

Model-centric metrics: Table II reports the faithfulness metrics, which are computed on the clip level. It appears that perturbation-based methods can reflect the model decision more faithfully, which could be explained by the fact that attributions are computed directly based on the output change; therefore, the energy of the heatmaps is more concentrated towards the real anomalies, on top of offering a better coverage

TABLE I: Results on human-centric metrics (unsupervised localization) over 40 clips. ↑ means higher is better, bold is best.

XAI methods	Clip level			Max-attribution-frame level		
	IoU↑	**F1↑**	**AUC↑**	**IoU↑**	**F1↑**	**AUC↑**
IG	0.1132	0.2014	0.6774	**0.1950**	**0.3186**	0.6962
IG+SmoothGrad	0.1008	0.1802	0.7134	0.1891	0.3039	**0.7501**
Guided BP	0.1231	0.2170	0.6243	0.1914	0.3171	0.6089
GradientSHAP	0.1115	0.1987	0.6769	0.1850	0.3047	0.6930
KernelSHAP	0.0837	0.1458	0.6746	0.1113	0.1870	0.6639
LIME	**0.1515**	**0.2433**	0.7259	0.1654	0.2604	0.7164
Occlusion_lp	0.0972	0.1684	**0.7432**	0.1306	0.2134	0.7707
Occlusion_sp	0.0986	0.1722	0.5814	0.1723	0.2609	0.6489

TABLE II: Results on model-centric metrics (faithfulness), robustness metric over 40 clips and run time per sample. ↑ / ↓ means higher/lower is better, bold is best.

XAI methods	Model-centric		Robustness			Runtime↓
	Deletion↑	**Insertion↓**	**mAUC↑**	**stdAUC↓**	C_v ↓	(m:s.ms)
IG	1.42E-01	3.17E-01	0.6821	4.61E-02	6.76E-02	00:10.7
IG+SmoothGrad	7.42E-02	3.76E-01	0.7254	5.65E-02	7.84E-02	00:55.7
Guided BP	3.05E-03	5.75E-01	0.6276	**4.06E-02**	**6.53E-02**	**00:00.5**
GradientSHAP	1.42E-01	3.03E-01	0.6817	4.52E-02	6.65E-02	00:02.0
KernelSHAP	1.92E-01	**9.91E-02**	0.6747	9.37E-02	1.55E-01	00:03.2
LIME	2.82E-01	1.41E-01	0.7257	9.10E-02	1.34E-01	00:03.2
Occlusion_lp	**2.99E-01**	1.46E-01	**0.7513**	8.14E-02	1.21E-01	00:07.9
Occlusion_sp	2.01E-01	1.01E-01	0.5868	1.44E-01	2.41E-01	10:23.2

of the whole objects than the other methods. The limitations of the gradient techniques described before (such as the partial coverage of the important objects) also leads to a decrease in insertion and deletion scores. Guided backpropagation is the worst method on either deletion or insertion metric. LIME and Occlusion with large perturbations perform significantly better on the deletion test.

Robustness metrics: The second part of Table II reports the mean and standard deviation of the previously obtained AUCs across the time axis, and the corresponding coefficient of variations (C_v). We observe that gradient methods are more stable than perturbation based ones, since gradient-based methods have a consistent localization behaviour regardless of their accuracy, whereas the localization of perturbation-based methods can be impacted by randomized initialization or sampling on perturbations. In particular, KernelSHAP is the least stable, and also has poor general anomaly localization.

Runtime: The overal computation cost largely depends on the number of required forward and backward passes. Guided backpropagation is the fastest method as a result of the single required back-propagation pass. Occlusion with small perturbations has the largest cost, since small windows results in a large number of computations (especially in the case of video data). Table II also shows that IG+SmoothGrad is not ideal, due to the large number of samplings for gradient integrals and noise-smoothing. This results in a trade-off when considering higher performance or smaller cost.

V. Conclusion

To comprehensively evaluate explanable AI methods in the context of video anomaly detection, we proposed an evaluation protocol based on used human-centric, model-centric and robustness metrics. We conclude that gradient-based methods generally achieve better robustness across multiple frames while perturbation methods reflect the model decision more faithfully with overall better insertion and deletion scores. From the user perspective, we found that LIME and IG+SmoothGrad outperform other methods on the localization of anomalous features, namely, at the clip level.

Acknowledgement

This research received funding from the FWO (Grant 1SB5721N), Belgium.

References

[1] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[2] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[3] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[4] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[6] E. Mahareek, E. El-Sayed, N. El-Desouky, and K. El-Dahshan, "Detecting anomalies in security cameras with 3DCNN and ConvLSTM," *preprint from Research Square*, 2023.

[7] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[8] L. Hiley, A. Preece, and Y. Hicks, "Explainable deep learning for video recognition tasks: A framework & recommendations," *arXiv preprint arXiv:1909.05667*, 2019.

[9] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, 2013.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[11] C. He, J. Shao, and J. Sun, "An anomaly-introduced learning method for abnormal event detection," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29 573–29 588, 2018.

[12] D. Koshti, S. Kamoji, N. Kalnad, S. Sreekumar, and S. Bhujbal, "Video anomaly detection using inflated 3D convolution network," in *2020 International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2020, pp. 729–733.

[13] S. Zhu, C. Chen, and W. Sultani, "Video anomaly detection for smart surveillance," in *Computer Vision: A Reference Guide*. Springer, 2020, pp. 1–8.

[14] X. Situ, I. Zukerman, C. Paris, S. Maruf, and G. Haffari, "Learning to explain: Generating stable explanations fast," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5340–5355.

[15] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *The Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010.

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

[17] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3319–3328.

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

[19] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[22] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[23] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI," *arXiv preprint arXiv:2201.08164*, 2022.

[24] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[25] G. F. Reed, F. Lynn, and B. D. Meade, "Use of coefficient of variation in assessing variability of quantitative assays," *Clinical and Vaccine Immunology*, vol. 9, no. 6, pp. 1235–1239, 2002.

[26] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," 2020.