Explanations for Computer Vision - XAI
RISE and LIME Explanations for Object Detection

Zhuoling Li
TU Munich
M.Sc. Informatics
zhuoling.li@tum.de

Sherif Nekkah
TU Munich
M.Sc. Robotics, Cog., Int.
sherif.nekkah@tum.de

Ishwor Subedi
TU Munich
M.Sc. Informatics
ishwor.subedi@tum.de

Abstract

While many state-of-the-art deep learning techniques have achieved remarkable performance in many domains, their incomprehensible decisions in terms of interpretability may in some sense limit their reliability in practical applications. Particularly in the vision domain, deep neural networks are mostly challenging for explainable artificial intelligence (XAI) that provides users with explainable decisions of those models. To enhance visual interpretations for these black-box models, this work proposes an object detection explainer that provides transparent explanations specifically for object detectors. It evaluates the contribution of each feature on the object detection results and thereby identifies the factors that mostly drive the detector's predictions, without accessing the internal structures of the detection models.

Keywords: XAI, explainable computer vision, object detection, LIME, RISE

1 Introduction

Understanding the internal processes in artificial neural networks is crucial to their performance and evaluation. Many aspects in a machine learning pipeline can cause a neural network to overfit and not generalize. Most sources of that problem lie in the used dataset. (Ying, 2019) Either not enough data or even small gaps due to edge cases can be fatal for the performance and generalizability of a neural network. Unfortunately, such inconsistencies, depending on the order of the problem, are very difficult to spot and require the appropriate set of tools.

Computer vision lately has experienced tremendous advances, mainly dominated by neural network architectures. Proposed models outperform each other almost on a monthly basis. Sometimes researchers have a hard time giving solid reasons why a proposed model works so well for their use case. Although most of the research doesn't go into production, at least until now, the examination of such models becomes more and more relevant. As autonomous systems are driven by artificial intelligence, such as driverless cars, the understanding of such models becomes crucial to their reliability, and ultimately to human safety. Unfortunately it is not straightforward to find explanations for a neural network model. Especially the evaluation of whether the explanation is qualitatively good or bad, is challenging. Some of the possible tools for explaining classifications and evaluation of these explanations were examined in this work.

We propose a concept, that aims at explaining object detections of any image-based object-detector model, using the explainers RISE and LIME. Mentioned explanations are further evaluated and compared using insertion and deletion as proposed by (Petsiuk et al., 2018).

2 Related work

In this section, we briefly introduce the state-of-the-arts in object classification and detection and enumerate recent research in explanations for computer vision.

2.1 Object Classification and Detection

Due to its wide variety of use-cases and applications, object-detection probably being the most common one, computer vision has seen vast advances in the last few decades. Especially with the increasing amount of publicly available open-source benchmark-datasets such as MS-COCO (Lin et al., 2014) and many more ((Everingham et al., 2010), (Deng et al., 2009), (Kuznetsova et al., 2018)), state-of-the-art algorithms and models are constantly improved.

Mainly due to the better exploitation of data, deep-learning-based approaches have taken the lead in image-based object detection. While neural network architectures allow countless possibilities for promising results, the advances in object detection were strongly marked by two works and their successors, being You Only Look Once (YOLO) (Redmon et al., 2015) and two-staged object-detection using Region-Based Convolutional Neural Networks (RCNN) (Girshick et al., 2013). While YOLO is framing object detection as a regression task as one big module, the structure of RCNN is divided into a region proposal module and a classifier. (Zaidi et al., 2021)

While object detectors vary greatly in their approach and structure, the underlying feature extraction module is crucial to achieving high performance. A feature extractor can be described as a model that maps an input image onto a set of high-level features or a latent space. This is also often referred to as a feature-encoder or model backbone and is of major interest for Explainable Artificial Intelligence (XAI). For further information on state-of-the-art object detectors and feature extractors that were used in this work, such as feature pyramid networks (FPN), please review (Zaidi et al., 2021).

2.2 Explanations in Computer Vision

With an increasing amount of applications in machine learning, recently a lot of questions are being raised about its authenticity. To tackle this, a lot of research has been going on in the field of explanations and the interpretability of machine learning models. As a result, multiple explanation methods help explain black-box models. LIME (Ribeiro et al., 2016) and RISE (Petsiuk et al., 2018) are two of the many available methods addressing this problem.

LIME Local Interpretable Model-agnostic Explanations is a method that explains single predictions of any classifier by locally learning an interpretable model around the prediction. One of the main advantages of using model agnostic explanations method like LIME is that it can be implemented regardless of the complexity of the underlying black-box model. The idea behind LIME is that black box models might be complex globally but they can be assumed to be linear locally. A linear surrogate model like LASSO (Tibshirani, 1996) is trained by perturbing the input and generating a dataset. The surrogate model is localized by weighting the perturbed samples by L2 distance. The coefficients of this linear model are then regarded as feature importance. In the case of images, superpixels are taken as features for the surrogate model and generated using any image segmentation algorithm, such as Quickshift as proposed by (Vedaldi and Soatto, 2008). LIME then perturbs the input image by rescaling the features of superpixels to zero or one. Zero means the superpixel is greyed out and one means that the superpixel is untouched.

RISE Randomized Input Sampling for Explanation (RISE), is a commonly used explanation method for image classification on single images. Similar to LIME, RISE does not require an understanding of the underlying black-box model. RISE, however, doesn't approximate a linear which might lead to more trustworthy results, especially for sufficiently complex models. (Petsiuk et al., 2018)

RISE is aiming to explain the importance of image regions for the classification of a particular class. This is done by observing the influence of perturbations on the model output. For that, RISE generates a random set of N binary masks that are scaled up to the size of the input image. Hereby the grid-size of the binary masks is a hyperparameter that determines the granularity of the explanation. Every mask is applied to the explained image individually and inputted to the black-box model. As the masks cover parts of the explained image, the probability score of the explained class is affected. Predictions with high scores will indicate that the important regions of the cover are not covered. A saliency map can be generated by a weighted sum of the probability scores of the perturbed images and the used masks, as shown in Figure 1.

Surrogate Object Detection Explainer (SODEx) (Sejr et al., 2021) is an abstract algorithm that proposed to provide explanations for black-box object detectors by any explainer. (Sejr et al., 2021) instantiated the algorithm by explaining YOLOv4 (Bochkovskiy et al., 2020) with LIME (Ribeiro et al., 2016). For each image, instead of directly explaining the object detection, SODEx explains a surrogate binary classifier that acquires a class score for the object under explanation, which corresponds to the object with the highest Intersection Over Union (IoU) score output by YOLOv4. The class score is derived from the probability score of YOLOv4 and indicates the confidence that the object under explanation is contained in the image. LIME will then measure the contribution of each (super)pixel on this class confidence, revealing the impact of each feature in the image on the resulting detection.

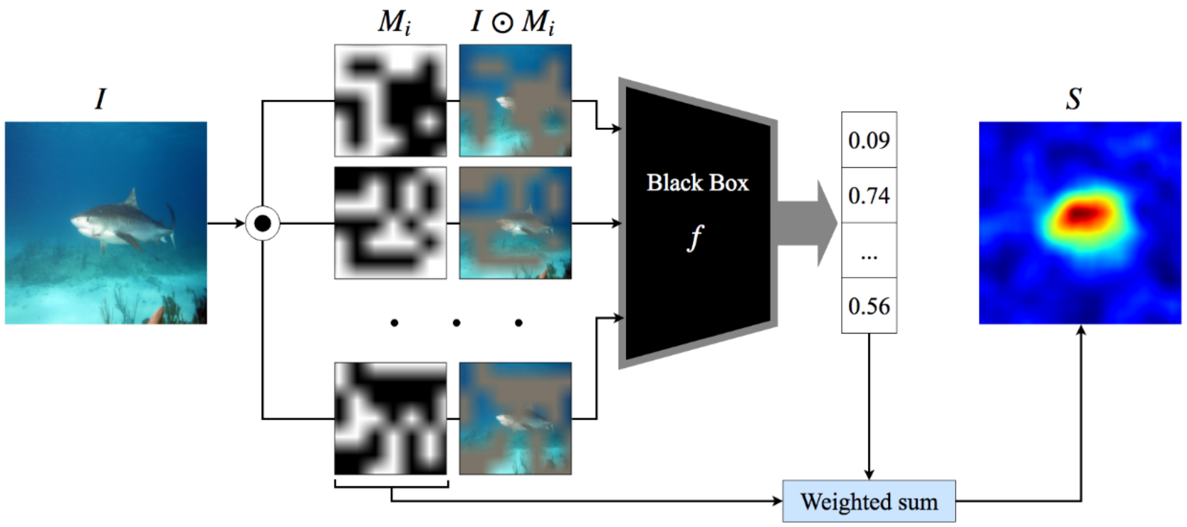


Figure 1: RISE explainer as proposed by (Petsiuk et al., 2018).

3 Explanations for Object Detection

The goal of our approach is to enrich the explainability of black-box object detection models. We proposed to extend LIME and RISE to object detection by leveraging the SODEx concept, where we instantiated object detector with Faster R-CNN. The overall pipeline is shown in Figure 2.

Object Detection and Selection For a given image, we aim to explain the object detection for one specific class, assuming that the objects labeled by the class are contained in the image. All detections together with their corresponding confidence scores will then be output by the object detector, where the most confident object (with the highest score) will represent the class under explanation. In other words, we explain the detection of the class by explaining the detection of this extremely confident object.

After determining the detection to be explained, the next step is to measure the contribution of each feature (in this case pixel or superpixel) of the image on this detection by the explainer. LIME and RISE explanations are differentiated here in terms of explaining strategies and discussed thus separately:

LIME Explanation LIME treats superpixels as the features of the image and therefore segments the image into superpixels by utilizing a segmentation algorithm, e.g., Quickshift (Vedaldi and Soatto, 2008). Next, we construct a binary classifier for the object under explanation, where the class probability corresponds to the score acquired by the object detector. LIME will further explain this binary classifier, i.e., LIME assigns features (superpixels) different weights which are regarded as feature importance, and indicate how much positive/negative impact each superpixel has on the detection.

RISE Explanation In our approach, RISE uses the instantiated object-detector as its black-box model in the general approach (Petsiuk et al., 2018). Instead of performing pure classifications, the model outputs a set of object detections with confidence scores for every perturbation. In every iteration, the highest score for the explained class is used to generate the saliency map. This further allows to generate a set of explanations for all classes, similarly using the highest corresponding score of detection for every perturbation.

4 Experiments

This section describes our experimental setup and presents quantitative and qualitative results.

4.1 Datasets

Penn-Fudan Dataset Since we are interested in explaining the detection of one given class, a dataset with one class is plausible in the initial stage, from which a simplified experimental procedure additionally benefits. Therefore, we introduced the Penn-Fudan Pedestrian Database (Wang et al., 2007) that contains 170 images with 345 labeled pedestrians. It concentrates only on the pedestrian class and assumes the presence of at least one pedestrian per image, and thus ensures that our model could provide explanations for at least one class initially.

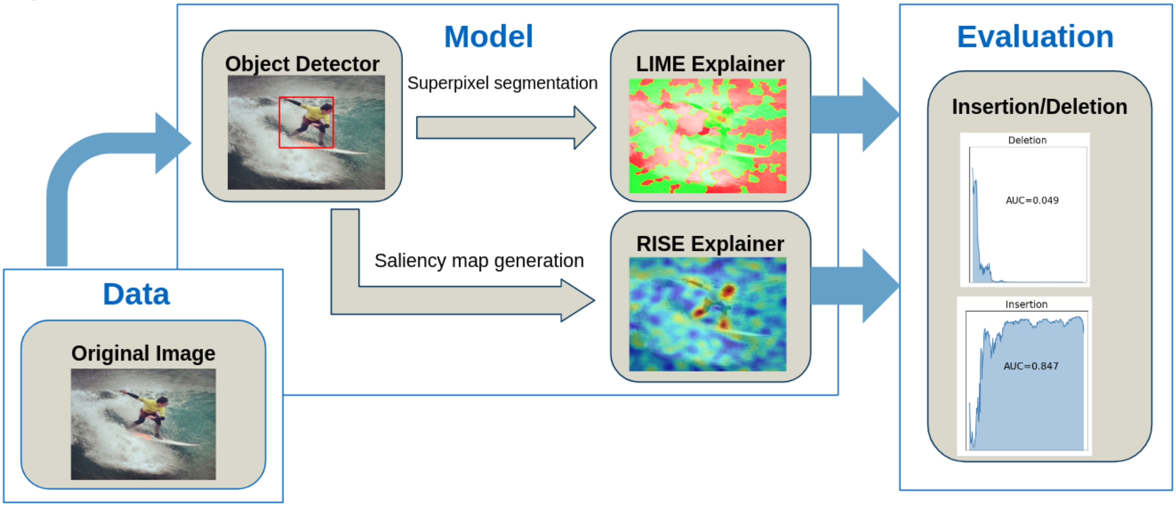


Figure 2: The pipeline of the proposed approach.

MS-COCO We would expect our model to generalize to more classes than just pedestrians. We've seen that deep learning models, particularly in the vision domain, have gained extraordinary performance with increasingly large datasets. MS-COCO (Lin et al., 2014) is established as a large-scale hand-labeled dataset contributing to different vision tasks such as object detection, segmentation, etc., with more than 200K labeled images and 91 categories included. Enriched object categories will promote a better generalization and avoid overfitting of the detector, and consequently facilitate a reasonable explanation later.

4.2 Implementation

Object Detector The main reasons to use Faster-RCNN are their dominance on vision-benchmark datasets such as COCO (Lin et al., 2014), their adaptability in incorporating state-of-the-art feature extractors as backbones, and their straightforward usability using the PyTorch and Torchvision libraries in python (Paszke et al., 2017). For the experiments on the Penn-Fudan dataset, a Faster-RCNN with a lightweight Resnet-18 backbone was used (He et al., 2015). For the training on MS-COCO, with 91 classes, a feature extractor with a lot more capacity was necessary. Therefore we switch from a Resnet-18 to a Resnet-50 FPN as a backbone.

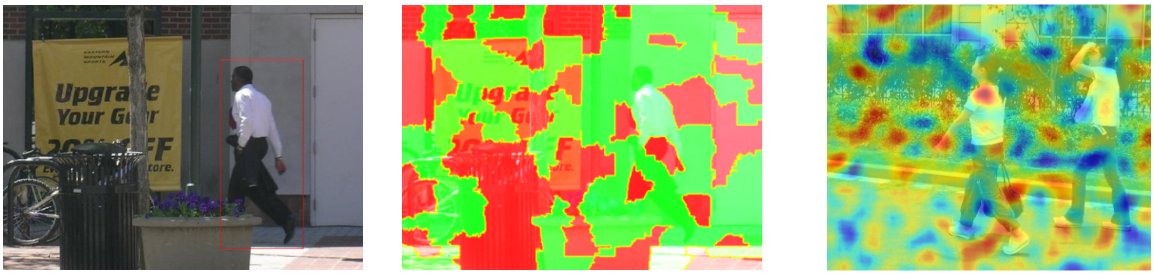
Hardware Initial experiments, being neural network training and a small amount of LIME and RISE explanations were run on Google Colab on consumer graded portable computers. For larger training, especially for the Faster-RCNN Resnet-50 FPN on COCO, this was unfeasible. To leverage the problems of limited storage and computing power, further training and explanations were performed on a Linux ppc64le platform with an NVIDIA Tesla v100 GPU with 32GB memory.

4.3 Evaluation Metrics

To evaluate the explanations quantitatively, we adopt the *insertion and deletion* metric, which was initially proposed in by (Petsiuk et al., 2018) for object classification. The idea of deletion is that if important pixels are deleted from the input image, it will influence the decision of the base model. In this case, there should be a sharp drop in the probability if the most important pixel in the image is removed. A low AUC score indicates the high efficiency of the model. Similarly in the insertion metric, if pixels are added according to their importance, there should be a steep rise in the probability, resulting in the high AUC score indicating a better model.

4.4 Results

Initially, we tested the proposed model on the Penn-Fudan dataset. The visual explanations of both LIME and RISE for the pedestrian detection together with the object detection are showed in Figure 3. We further conducted the experiment on the MS-COCO dataset. Here a sample explanation on the surfer object is generated in Figure 4. Additionally, the evaluation with insertion/deletion metric on RISE explanation of Figure 4c is visualized in Figure 6.

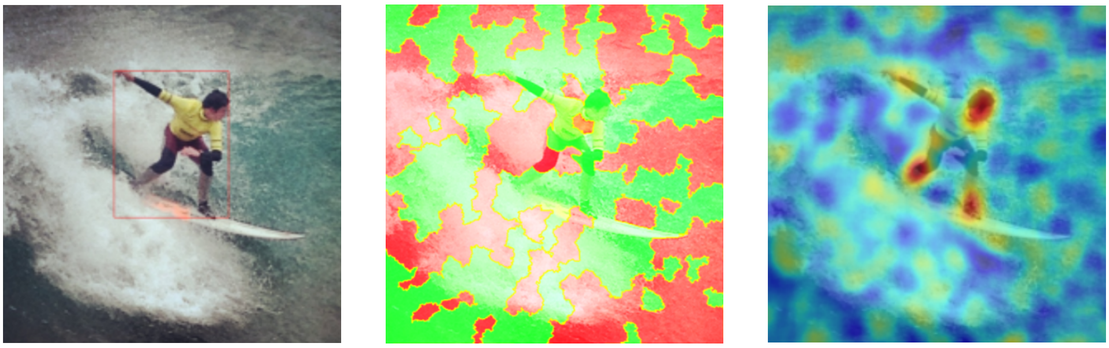


(a) Object detection.

(b) LIME explanation.

(c) RISE explanation.

Figure 3: Results on the Penn-Fudan dataset.

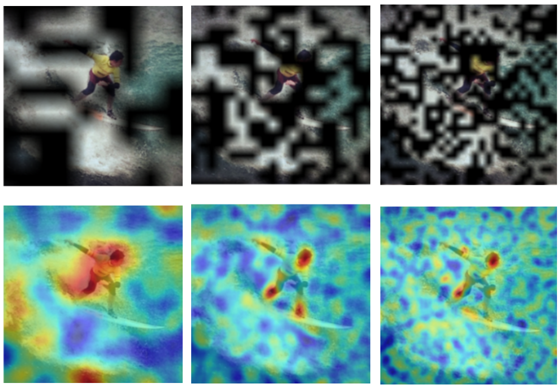


(a) Object detection.

(b) LIME explanation.

(c) RISE explanation.

Figure 4: Results on the MS-COCO dataset.



(a) $s = 10$

(b) $s = 20$

(c) $s = 30$

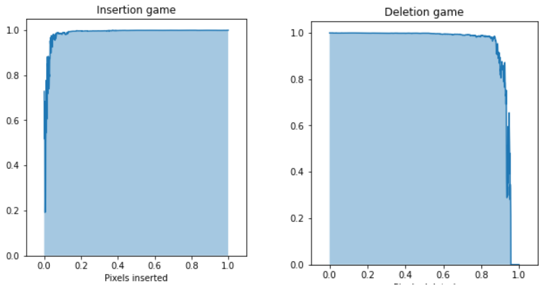
Figure 5: RISE explanations on MS-COCO dataset with different grid-sizes s for the binary masks. Up: Exemplary binary masks, Down: Explanations.

4.5 Analysis

Qualitative Analysis The visualized results on the two datasets both pointed out that RISE explanations are more comprehensible compared to LIME explanations. Most dominant pixels locate within the bounding box of the detected object while LIME explainer also takes disperse superpixels outside of the bounding box as positive features for the explanations. Moreover, the comparison between Figure 3 and Figure 4 evidently demonstrated pretraining on large-scale datasets will promote the learning of the detector and the subsequent explanations.

Parameter Sensitivity Figure 5 shows three different RISE explanations of the same object detection. The difference between the three explanations is that the binary masks, that are generated for the perturbations, have different grid sizes. Intuitively, the second explanation, with a grid size of $s = 20$, as shown in Figure 5b is the best, as it is a trade-off between a low number of artifacts and a high explanation quality. These explanations were generated with a number of $N = 1000$ samples. In theory, the explanations with the smallest grid size should yield the most accurate explanation. In practice, however, as computing power is limited and the number of samples is restricted, a wise choice of hyperparameters is necessary.

Quantitative Analysis As the AUC curves illustrated in Figure 6, a steep rise of the probability at the beginning of the insertion is as expected, resulting a high AUC score. That implies that the object detector works efficiently, i.e., it mainly concentrates on the pixels that are crucial to the detections. On the other hand, a sharp drop of the curve does not exist when we delete the more important pixels. It might be potentially caused by the fact that dif-



(a) Insertion.

(b) Deletion

Figure 6: Evaluation of RISE explanation with insertion and deletion metric

ferent from image classification, object detection is much more sensitive to some important pixels (e.g., the pixels that shape the objects). Therefore, deleting important pixels would mislead the detector's decisions and therefore make the deletion metric ineffective.

5 Conclusion

We presented an object detection explainer that provides an interpretable output of any object detector, utilizing LIME and RISE explanations. Due to performance and practical reasons we instantiated the detector with Faster R-CNN in this work. The model measures the contribution of each feature in the image on the final object detection, representing different feature importance. For evaluation, we experimented with the model on both the Penn-Fudan and the MS-COCO dataset with different network architecture details. The pedestrian dataset allowed the use of the lightweight feature extractor ResNet-18. The MS-COCO dataset demanded way more capacity, thus a ResNet-50 FPN was used. The resulting explanations transparently showed that pixels inside a bounding box of an explained detection tend to contribute more to classification than pixels outside of the bounding box. The visualizations demonstrated that RISE explanations are more interpretable than LIME explanations, as very specific fine-grained features are contributing towards the detection of a certain class. Furthermore, we have shown that the explainers themselves also rely on a set of hyperparameters and that the resulting explanations are no factual pieces of information. Therefore the evaluation of the explanations must be considered.

6 Outlook

Our experimental results clearly showed that RISE achieved a better performance than LIME. One underlying cause might be that RISE perturbs the images in an "average manner", i.e., the saliency is computed by the weighted sum of a set of generated masks, while LIME's perturbation is relatively more random.

Therefore, we envision improving the model in a way that we may also consider leveraging the average concept in LIME, e.g., we could perturb the images multiple times and average the resulting weights, in order to eliminate bad cases caused by randomness. Additionally, it's worth mentioning that the authors of RISE also took this idea into account and proposed recently another saliency-based explainer called D-RISE (Petsiuk et al., 2021) that also extended from RISE for detection tasks, where they generate multiple saliency maps and take the average. The remarkable results described in this paper also justified that eliminating randomness can significantly improve the efficiency of the explainer.

References

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models.

Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. 2021. Black-box explanation of object detectors via saliency maps.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2015. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Jonas Herskind Sejr, Peter Schneider-Kamp, and Naeem Ayoub. 2021. Surrogate object detection explainer (sodex) with yolov4 and lime. *Machine Learning and Knowledge Extraction*, 3(3):662–671.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Andrea Vedaldi and Stefano Soatto. 2008. Quick shift and kernel methods for mode seeking. In *Computer Vision – ECCV 2008*, pages 705–718, Berlin, Heidelberg. Springer Berlin Heidelberg.

Liming Wang, Jianbo Shi, Gang Song, I-fan Shen, et al. 2007. Object detection combining recognition and segmentation. In *Asian conference on computer vision*, pages 189–199. Springer.

Xue Ying. 2019. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168:022022.

Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Naveed Asghar, and Brian Lee. 2021. A survey of modern deep learning based object detection models. *CoRR*, abs/2104.11892.