









# Explainable Artificial Intelligence for Object Detection in the Automotive Sector <sup>†</sup>

Marios Siganos <sup>1</sup>, Panagiotis Radoglou-Grammatikis <sup>1,2,\*</sup>, Thomas Lagkas <sup>3</sup>, Vasileios Argyriou <sup>4</sup>,  
Sotirios Goudos <sup>5</sup>, Konstantinos E. Psannis <sup>6</sup>, Konstantinos-Filippos Kollias <sup>2</sup>, George F. Fragulis <sup>2</sup>  
and Panagiotis Sarigiannidis <sup>2</sup>

<sup>1</sup> K3Y Ltd., Studentski District, Vitosha Quarter, Bl. 9, 1700 Sofia, Bulgaria; msiganos@k3y.bg

<sup>2</sup> Department of Electrical and Computer Engineering, University of Western Macedonia, Campus ZEP Kozani, 50100 Kozani, Greece; dece00063@uowm.gr (K.-F.K.); gfragulis@uowm.gr (G.F.F.); psarigiannidis@uowm.gr (P.S.)

<sup>3</sup> Department of Informatics, Democritus University of Thrace, Kavala Campus, 65404 Kavala, Greece; tlagkas@cs.duth.gr

<sup>4</sup> Department of Networks and Digital Media, Kingston University London, Penrhyn Road, Kingston upon Thames, Surrey KT1 2EE, UK; vasileios.argyriou@kingston.ac.uk

<sup>5</sup> School of Physics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; sgoudo@physics.auth.gr

<sup>6</sup> Department of Applied Informatics, School of Information Sciences, University of Macedonia, 156 Egnatia Street, 54636 Thessaloniki, Greece; kpsannis@uom.edu.gr

\* Correspondence: pradoglou@k3y.bg

<sup>†</sup> Presented at the 7th International Global Conference Series on ICT Integration in Technical Education & Smart Society, Aizuwakamatsu City, Japan, 20–26 January 2025.

## Abstract

In the automotive domain, object detection is pivotal for enhancing safety and autonomy through the identification of various objects of interest. However, insights into the influential image pixels in the detection process are often lacking. Recognizing these significant regions within the image not only enriches our qualitative understanding of the model's functionality but also empowers us to refine and optimize its performance. Employing Explainable Artificial Intelligence (XAI), we present an XAI component in this paper. This component explains the predictions made by a pre-trained object detection model for a given image by generating heatmaps that highlight the most critical regions in the image for the detected objects.

**Keywords:** artificial intelligence; automotive; EigenCAM; explainable ai; explainable artificial intelligence; object detection; visual XAI; XAI; YOLO



Academic Editors: Debopriyo Roy and Peter Ilic

Published: 1 September 2025

**Citation:** Siganos, M.; Radoglou-Grammatikis, P.; Lagkas, T.; Argyriou, V.; Goudos, S.; Psannis, K.E.; Kollias, K.-F.; Fragulis, G.F.; Sarigiannidis, P. Explainable Artificial Intelligence for Object Detection in the Automotive Sector. *Eng. Proc.* **2025**, *107*, 44. <https://doi.org/10.3390/engproc2025107044>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In automotive applications, object detection plays a critical role in enhancing safety and autonomy by identifying and classifying various objects in the vehicle's surroundings, such as pedestrians, vehicles, cyclists, and road signs. Explainable Artificial Intelligence (XAI) techniques can be employed to provide insights into the decision-making process of these object detection systems, offering explanations for the detections made.

In the automotive industry, XAI is very important due to the high stakes involved in autonomous driving applications. Users and regulators demand transparency and accountability from autonomous systems, and XAI helps fulfill these requirements by providing interpretable insights into the decision-making processes of object detection algorithms.

An example is the case where an autonomous vehicle identifies a pedestrian crossing the road. XAI can provide explanations for why the pedestrian was detected, which features

in the image led to the detection, and how confident the system is in its decision. These explanations not only enhance the transparency of the system's behavior but also facilitate trust and understanding among users, regulators, and stakeholders.

This paper focuses on the application of XAI in the context of object detection in the automotive domain. More specifically, it presents a software component that involves explainability methods for object detection systems deployed in vehicles. This component can serve as an explainability module within a broader framework or system comprising multiple stages including pre-processing, detection, explainability, and notification, like the one proposed in [1].

The paper is structured as follows: Section 2 provides a review of related literature. Section 3 offers background information on relevant terms and approaches. In Section 4, the detailed architecture of the XAI component is presented. Section 5 presents the experimental findings. Finally, Section 6 concludes the paper with reflections on lessons learned and outlines future directions.

## 2. Related Work

Numerous studies have explored XAI applications within the automotive sector. These encompass diverse tasks such as object detection, predicting dangerous vehicle behaviors, recognizing pedestrian intentions, identifying traffic lights, estimating steering angles, and classifying images concerning traffic signs and vehicle/non-vehicle distinctions. To explain the predictions of the proposed ML methods, researchers employ various XAI techniques including saliency maps, Grad-CAM, sensitivity analysis, and visual attention mechanisms. These approaches primarily generate heatmaps to highlight significant image regions.

Mankodiya et al. [2] introduce a semantic object detection method for autonomous vehicles based on XAI. The study involves training and comparing three deep learning architectures—ResNet-18, ResNet-50, and SegNet—for road detection. Various XAI methods are employed to explain the predictions of these black-box models. The best-performing model's predictions are specifically explained using the Grad-CAM and saliency map XAI techniques, which generate visual heatmaps highlighting important regions in the images. For training and testing, the KITTI road dataset, comprising 289 road images and their corresponding annotations, is utilized. The results indicate that ResNet-18 exhibits superior prediction performance, and the brighter pixel values in the generated heatmap highlight more significant regions in the image. Despite these insights, the authors conclude that while the explanations provide useful insights, they do not fully address the challenges associated with black-box interpretability.

The authors in [3] propose a comprehensive vision-based framework tailored for autonomous driving, encompassing four critical tasks: object detection, dangerous vehicle prediction, pedestrian intention recognition, and traffic light identification. They leverage various datasets to train and validate their models, including the BDD100K dataset for object detection, a pedestrian dataset for intention prediction, a vehicle dataset for dangerous vehicle estimation, and a traffic light dataset for signal recognition. In the object detection task, the model identifies various objects within the driving environment, providing bounding boxes and probabilities for each category. For pedestrian intention recognition, the framework utilizes YOLOv4 to detect pedestrians and subsequently employs Part Affinity Fields (PAFs) to extract human skeleton features. A Convolutional Neural Network (CNN) then analyzes these features to infer pedestrian intentions, with results displayed as labels on bounding boxes. Dangerous vehicle prediction involves classifying vehicles as safe or dangerous based on behaviors such as braking, turning left, turning right, or crossing. This classification is achieved through a CNN model, with the fine-tuned EfficientNet model demonstrating superior accuracy compared to other examined models, such as MobileNet,

VGGNet, GoogLeNet, and ResNet. Similarly, for traffic light recognition, the model distinguishes between safe signals (e.g., green lights) and dangerous signals (e.g., red lights). Experimental results reveal that the fine-tuned MobileNet model outperforms alternative models. To enhance the interpretability of their models, especially in the risk estimation phase involving CNN-based models for vehicle and traffic light identification, the authors apply the Randomized Input Sampling for Explanation (RISE) algorithm. This technique generates saliency maps, revealing the importance of each pixel in determining the final classification results.

In [4], a novel XAI method tailored for CNN-based models utilized in self-driving cars is introduced. The method primarily focuses on sensitivity analysis, which gauges the impact of each input feature. Initially, the images are converted to grayscale and are subsequently altered from their original form. These modified and original images are then fed into the model, and their disparities are evaluated by comparing the resulting output vectors. Then, an explanation is generated by proportionally comparing each segment of the original image to the previously computed difference, effectively pinpointing the influential sections contributing to the final prediction. The efficacy of this XAI approach is demonstrated through its application on a pre-trained CNN across four distinct image datasets containing vehicle/non-vehicle images and images featuring traffic signs. Each prediction is explained using this method, highlighting every step of the algorithm. These explanations are then compared to those produced by other XAI techniques such as SHAP, LIME, Grad-CAM, and eXplainable CNN (XCNN). Notably, the proposed method consistently delivers successful explanations for all images, unlike other methods that may produce inadequate explanations.

In [5], a framework aimed at explaining decisions made in autonomous driving scenarios through visual attention mechanisms is introduced. To tackle this challenge, the authors adopt an Imitation Learning approach, wherein a driving policy is acquired from RGB frames to associate observed frames with corresponding vehicle steering angles. The framework employs a fully convolutional network comprising five layers to extract feature maps from input images. These feature maps are then fed into a multi-head visual attention block to explain the predictions, followed by a final prediction block that estimates the steering angle, representing the intended driver action. The model's effectiveness is demonstrated utilizing data from the CARLA urban driving simulator. The visual attention activations are presented, showcasing how specific regions within the image are weighted to offer insights into prediction explanations. Results indicate that integrating visual attention mechanisms not only enhances prediction explanations but also boosts overall model performance.

### 3. Background

#### 3.1. Computer Vision and Object Detection

Computer vision is a field of Artificial Intelligence (AI) that focuses on enabling computers to interpret and understand visual information from the world around them. It encompasses a wide range of tasks, including image classification, object detection, image segmentation, and image generation. This paper focuses on object detection, which involves identifying and locating objects of interest within an image or video. This task is crucial for various applications, including autonomous vehicles, surveillance systems, medical imaging, and augmented reality. Object detection algorithms analyze input images or video frames to detect and localize objects by predicting bounding boxes around them and assigning class labels to the detected objects.

There are different approaches to object detection, including one-stage and two-stage algorithms, each with its own trade-offs between speed and accuracy. One-stage detection

algorithms, such as YOLO (You Only Look Once) [6] and SSD (Single Shot Multibox Detector) [7], are designed to directly predict bounding boxes and class probabilities in a single pass through the neural network. On the other hand, two-stage detection algorithms, like Faster R-CNN (Region-based Convolutional Neural Network) [8], consist of two sequential stages. In the first stage the algorithm generates region proposals, while in the second stage, the proposed regions are classified into different object categories. One-stage detection algorithms are known for their efficiency and speed, making them suitable for real-time applications. However, two-stage algorithms generally offer higher accuracy but can be computationally more intensive and slower compared to one-stage algorithms.

### 3.2. Explainable AI and Explainability

Explainable AI (XAI) refers to the development of AI systems that can provide understandable explanations for their decisions and behaviors. It aims to bridge the gap between the opaque nature of complex machine learning models and the need for transparency, accountability, and trust in AI systems. In the context of XAI, explainability refers to the ability of an AI system to provide insights into why specific decisions were made, how the system arrived at those decisions, and which factors influenced its output.

There are various approaches to achieving explainability in AI systems. First there are interpretable models that are simple models that humans can easily understand and interpret, such as decision trees or linear regression. Then there are transparent algorithms with built-in mechanisms for providing explanations, such as rule-based systems or probabilistic graphical models. Lastly, there are post hoc explanations, which involve the generation of explanations for decisions made by complex black-box models like deep neural networks, using techniques such as feature importance scores, attention maps, and sensitivity analysis.

Overall, explainability mechanisms are crucial for ensuring that AI systems are not only accurate and efficient but also trustworthy and accountable. By enabling humans to understand AI decisions, XAI promotes ethical AI development and deployment, ultimately leading to more responsible and beneficial use of artificial intelligence across various domains.

### 3.3. XAI for Images and Explainable Object Detection

In the context of images, XAI techniques aim to explain the inner workings of complex deep learning models, which often operate as black boxes. By employing methods such as feature visualization, saliency mapping, and attention mechanisms, XAI offers insights into which parts of an image influenced the model's decisions and why. This paper focuses on object detection where XAI techniques can produce explanations in the form of heatmaps, clarifying which image features led to the detection of specific objects and how confident the model is in its predictions. In general, the application of XAI to images aims to enhance the interpretability, trustworthiness, and accountability of AI systems, enabling users to better understand their decisions.

The Class Activation Maps (CAMs) proposed in [9] are a powerful technique that can be employed in object detection to provide insights into the regions of an image that contribute most significantly to the model's predictions. CAMs offer a visual representation of the discriminative regions within an image that influence the classification decision for a particular object class. They are useful in interpreting decisions of deep learning models, especially CNNs used in object detection tasks. They generate heatmaps that highlight the regions of interest, aiding users in visualizing model attention within the image.

Methods that can generate CAMs are divided into gradient-based methods like Gradient-weighted Class Activation Mapping or simply Grad-CAM and gradient-free

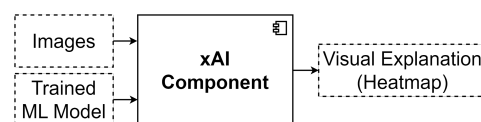
methods like EigenCAM. On the one hand, Grad-CAM generates CAMs by computing the gradients of the target class score with respect to the feature maps of the last convolutional layer [10]. By weighting these gradients, Grad-CAM highlights the importance of each feature map for the target class prediction. On the other hand, the EigenCAM method leverages the principal components derived from learned representations within the convolutional layers to generate visual explanations [11].

#### 4. Architecture

The proposed XAI component serves to explain the decision-making process of a pre-trained object detection model, such as YOLO, by providing interpretable explanations in the form of heatmaps. When presented with an input image and a pre-trained YOLO model, the XAI component first passes the image through the YOLO model for object detection. YOLO identifies and localizes objects within the image, providing bounding boxes and class probabilities for each detected object. Next, the XAI component extracts feature maps from various layers of the YOLO model that correspond to the detected objects. These feature maps encapsulate the activations and responses of the neural network to different regions of the input image. Subsequently, the XAI component employs EigenCAM, which utilizes the extracted feature maps to generate a heatmap that highlights the regions of the input image that contributed most significantly to the model's decision for each detected object class. EigenCAM has been selected as the preferred method because of its advantages over other state-of-the-art methods. Firstly, it seamlessly integrates with all CNN models, eliminating the need for modifications to layers or model retraining. Secondly, it can generate visual explanations for multiple objects within the same image. Lastly, it operates independently of gradient back-propagation, further enhancing its versatility and efficiency.

For the implementation of the XAI component, python was used, along with pytorch and the pytorch-gradcam library [12] version 2.8.0, which offers EigenCAM.

Figure 1 presents the high-level architecture of the proposed XAI component that takes images and a pre-trained model as input and generates a visual explanation in the form of a heatmap.



**Figure 1.** XAI component architecture.

The heatmap produced by the XAI component serves as an explanation, indicating which parts of the image were crucial in the YOLO model's classification of each object. Regions colored in red signify areas of high importance, indicating that these regions in the input image strongly influenced the model's prediction. Overall, the proposed XAI component enhances the transparency and interpretability of the YOLO model's predictions by providing intuitive visual explanations in the form of heatmaps, enabling users to understand and trust the model's decisions more effectively.

#### 5. Experimental Results

This section presents the results of the evaluation of the XAI component applied in the context of the automotive domain. For the evaluation of the XAI component, YOLOv5 [13] was employed as the input object detection model. An image depicting a road scene and featuring various elements such as a road, a car, a cyclist, a traffic light, and several traffic signs was passed as the input image into the XAI component.

The class activation maps for the pre-trained object detection model are shown in Figure 2. The first image is the original image passed as input. Detected objects in this scene



include a car, a cyclist, and a traffic light. In the second image, prominent regions are highlighted, notably the traffic light and multiple traffic signs. Lastly, the third image focuses on pinpointing areas that have the most influence on each individually detected object.



**Figure 2.** An illustration showcasing an original input image processed by the XAI component, alongside the output heatmap highlighting significant areas within the entire image, as well as individual heatmaps highlighting detected objects within the scene.

## 6. Conclusions

In this paper, a software component that focuses on the explainability of pre-trained object detection models in the automotive domain was presented. The proposed XAI component was evaluated by employing a YOLOv5 detection model to generate visual explanations for images related to the automotive domain. As part of the next steps and future directions, we plan to incorporate and evaluate additional CAM-based XAI methods and extend the assessment of the XAI component utilizing other state-of-the-art object detection models, such as Faster R-CNN and SSD. Lastly, we intend to assess the effectiveness of our XAI component through user satisfaction surveys and various quantitative metrics.

**Author Contributions:** Conceptualization, M.S. and P.R.-G.; methodology, M.S., P.R.-G., T.L. and V.A.; software, M.S., K.E.P., K.-F.K. and G.F.F.; validation, T.L., V.A. and P.S.; writing—original draft preparation, M.S., P.R.-G., T.L. and V.A.; writing—review and editing, S.G., K.E.P., K.-F.K. and G.F.F.; supervision, P.R.-G., T.L. and P.S.; project administration, P.R.-G.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No. 101070214 (TRUSTEE).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** No new data were created or analyzed in this study. Data sharing is not applicable to this article.

**Conflicts of Interest:** Marios Siganos and Panagiotis Radoglou-Grammatikis were employed by the company K3Y Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Siganos, M.; Radoglou-Grammatikis, P.; Kotsiuba, I.; Markakis, E.; Moscholios, I.; Goudos, S.; Sarigiannidis, P. Explainable AI-based Intrusion Detection in the Internet of Things. In Proceedings of the 18th International Conference on Availability, Reliability and Security, Benevento, Italy, 29 August–1 September 2023; pp. 1–10. [\[CrossRef\]](#)
2. Mankodiya, H.; Jadav, D.; Gupta, R.; Tanwar, S.; Hong, W.C.; Sharma, R. OD-XAI: Explainable AI-Based Semantic Object Detection for Autonomous Vehicles. *Appl. Sci.* **2022**, *12*, 5310. [\[CrossRef\]](#)
3. Li, Y.; Wang, H.; Dang, L.M.; Nguyen, T.N.; Han, D.; Lee, A.; Jang, I.; Moon, H. A Deep Learning-Based Hybrid Framework for Object Detection and Recognition in Autonomous Driving. *IEEE Access* **2020**, *8*, 194228–194239. [\[CrossRef\]](#)

4. Kim, H.S.; Joe, I. An XAI method for convolutional neural networks in self-driving cars. *PLoS ONE* **2022**, *17*, e0267282. [[CrossRef](#)] [[PubMed](#)]
5. Cultrera, L.; Seidenari, L.; Becattini, F.; Pala, P.; Bimbo, A.D. Explaining Autonomous Driving by Learning End-to-End Visual Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Seattle, WA, USA, 14–19 June 2020. [[CrossRef](#)]
6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
7. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [[CrossRef](#)] [[PubMed](#)]
9. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
10. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
11. Muhammad, M.B.; Yeasin, M. Eigen-cam: Class activation map using principal components. In Proceedings of the 2020 IEEE International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–7.
12. Gildenblat, J. PyTorch Library for CAM Methods. Available online: <https://github.com/jacobgil/pytorch-grad-cam> (accessed on 4 November 2024).
13. Jocher, G. YOLOv5 by Ultralytics. 2020. Available online: <https://zenodo.org/records/7347926> (accessed on 28 October 2024).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.