Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2026) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

30th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2026)

RF-BoxCAM: Receptive Field Guided High Resolution Visual Explanations for Object Detection Networks

Malek BEN HASSINE[a,c,b], Reza BAHMANYAR[a], Houda CHAABOUNI-CHOUAYAKH[b]

[a] *Remote Sensing Technology Institute, German Aerospace Center (DLR),Wessling ,Germany*
[b] *Sm@rts Laboratory, Digital Research Center of Sfax, Tunisia*
[c] *Higher School of Communication of Tunis (SUP'COM), Tunisia*

Abstract

Aerial object detection is a fundamental component of safety-critical applications such as surveillance, traffic monitoring, and autonomous systems. Although deep learning (DL) based object detectors have achieved strong performance, their lack of interpretability remains a major barrier to trustworthy deployment. This limitation is particularly pronounced in aerial imagery, where objects are often small, densely packed, and visually ambiguous. Existing explainable artificial intelligence (XAI) techniques, both black and white-boxes, often fail to provide meaningful and faithful explanations for small-object detections, as their attribution mechanisms tend to focus on coarse or irrelevant regions. In this paper, we analyze these shortcomings and propose Receptive Field guided Bounding Box Class Activation Map (RF-BoxCAM), a novel XAI method that is specifically designed and adapted for commonly used YOLO object detection architectures and specifically designed to address the challenges of small-object detection. The proposed approach produces efficient, object-centric visual explanations in a single forward pass, improving transparency while significantly reducing computational overhead compared to black-box methods. By enhancing interpretability for small-object aerial detection, this work contributes toward safer and more reliable deployment of DL models in safety-sensitive systems.

Keywords: Object Detection; Explainable AI; Aerial Imagery; YOLO; RF-BoxCAM

1. Introduction

Object detection in aerial imagery is a core component of many safety-critical systems, including traffic monitoring, disaster response, urban analysis, and autonomous aerial platforms. Recent advances in DL have led to the devel-

* Malek BEN HASSINE. Tel.: +49016096435685
E-mail address: malek.benhassine@supcom.tn

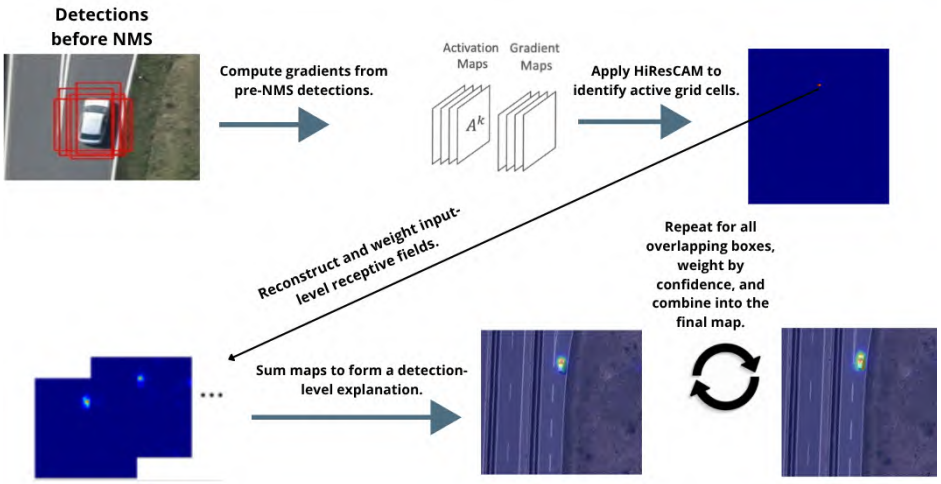

Fig. 1: Overview of the RF-BoxCAM pipeline: Pre-NMS detections are first evaluated using HiResCAM weighting to identify active grid cells. For each active cell, its input-level receptive field is reconstructed and weighted by importance. These individual maps are then aggregated across all bounding boxes targeting the same object to generate the final, object-centric saliency map.

opment of models that have achieved substantial improvements in detection accuracy on large-scale aerial image datasets. However, contemporary object detectors operate as complex, highly non-linear systems, producing predictions through deep hierarchies of convolutional and feature-fusion layers. This opacity makes it difficult to validate model behavior, diagnose systematic errors, or assess reliability under novel or adverse conditions [9]. Indeed, the lack of interpretability in object detection models remains a major barrier to their trustworthy deployment in real-world applications, where detection outputs are often used to inform decisions with significant consequences [18]. Understanding why a detector identifies a particular object with a certain confidence, or struggles to accurately localize it, is therefore as important as achieving high detection performance. These challenges are particularly pronounced in aerial imagery. Objects of interest, such as vehicles, are often small relative to the image resolution and densely distributed across the scene. They appear in visually complex environments containing repetitive textures, strong shadows, and background structures that resemble object features. Under such conditions, faithful explanations require precise spatial localization, as even minor attribution errors can shift focus away from the object and toward surrounding context. Explainable artificial intelligence (XAI) has emerged as a promising direction for improving the transparency of DL models [1]. A large body of prior work has focused on generating visual explanations for image classification models, where a single global prediction is produced for each input image. In this setting, two main families of explanation methods are commonly used: white-box techniques, which attribute importance through computing derivatives of the output with respect to feature maps, and black-box techniques, which assess relevance by observing prediction changes under controlled input modifications. Both approaches have demonstrated the ability to provide intuitive and informative explanations, with representative methods including Grad-CAM [4], HiResCAM [5], RISE[10], and more recent faithfulness-oriented attribution techniques [22]. In contrast, explainability for object detection remains less explored and inherently more challenging. Object detectors produce multiple predictions per image, each associated with spatial localization, class probabilities, and confidence scores. This multi-instance output structure introduces complex interactions between object appearance, spatial context, and scale, making it difficult to directly transfer classification-oriented explanation techniques to detection settings. As a result, many existing XAI methods fail to adequately capture the reasoning process of object detectors [2], particularly in scenarios involving dense scenes and small objects. These limitations motivate the need for explanation approaches that are specifically designed to address the structural properties of detection models. Therefore, in this work, we:

- adapted and assessed existing XAI methods for object detection in aerial imagery, identifying their limitations.
- propose **RF-BoxCAM**, a computationally efficient white-box XAI technique tailored for YOLO architectures that resolves conventional spatial inaccuracies for small object detection in aerial imagery.

- demonstrate the applicability of **RF-BoxCAM** on vehicle detection in an aerial image dataset, highlighting its potential to improve interpretability, reliability, and computational efficiency.

2. Related Work

XAI has been extensively studied in the context of image classification, where models generate a single prediction for each image. Existing approaches can be categorized as either white-box or black-box. White-box, techniques compute the sensitivity of the output with respect to input pixels or intermediate feature maps. Early saliency methods directly used input gradients to highlight influential regions [3]. Class activation mapping approaches such as Grad-CAM [4] project gradients onto convolutional feature maps to generate class-discriminative localization maps. HiResCAM [5] further improves spatial faithfulness by performing element-wise multiplication between activations and gradients before aggregation. These methods are computationally efficient and require only a single backward pass. Black-box, techniques, in contrast, estimate feature importance by measuring changes in predictions under systematic input modifications. LIME [7] trains local surrogate models to approximate predictions and SHAP [8] leverages cooperative game theory to assign feature importance, while RISE generates saliency maps through randomized masking of input images. These techniques are model-agnostic and often more faithful, but require a large number of forward passes, leading to substantial computational overhead.

While the previously cited techniques are effective for classification, they are fundamentally designed for single-label, global predictions and do not explicitly address localization or multi-instance outputs. Compared to classification, object detection poses additional challenges since models are supposed to localize objects by generating multiple predictions per image, each of which is associated with bounding box coordinates and class scores. White-box approaches adapt gradient-based techniques to detection models. Grad-CAM [4] and HiResCAM [5] can be applied to detection heads by selecting class confidence scores as explanation targets. These methods retain computational efficiency, as explanations can be generated with a single backward pass. However, when applied to modern multi-scale detectors with feature pyramid architectures, naive upsampling of feature-space attributions often produces spatially coarse or misaligned saliency maps. This limitation becomes particularly pronounced when explaining small objects, where precise spatial alignment is critical [6]. Black-box approaches extend perturbation strategies to the detection setting. Methods such as D-RISE [11] evaluate how masked regions influence detection confidence and localization accuracy. D-CLOSE [21] further refines this idea by modeling spatial consistency between perturbations and predicted boxes. These approaches often produce faithful, object-aware explanations but require hundreds or thousands of forward passes per detection, making them computationally expensive in dense scenes.

Existing XAI methods present a trade-off. Black-box approaches are faithful but computationally expensive for large-scale or real-time object detection, while white-box methods are efficient but rely on assumptions about spatial correspondence between feature maps and input pixels that may not hold in modern multi-scale detectors. These limitations motivate the development of XAI methods that remain efficient while explicitly considering the receptive-field structure of object detection architectures. In this work, we address these challenges by analyzing the failure modes of existing XAI techniques for detecting small objects in aerial imagery and by proposing a novel receptive-field-based explanation approach named **RF-BoxCAM**. By explicitly reconstructing the contribution of detector receptive fields at the input-image level, the proposed method produces localized, object-centric explanations that better align with the detector's internal reasoning. The approach is designed to be computationally efficient and applicable across different detection architectures, making it suitable for dense aerial scenes and safety-critical perception systems.

3. Methodology

This section details our proposed **RF-BoxCAM** methodology, a white-box explanation technique for modern multi-scale detectors. The full pipeline, from raw pre-NMS detections to the final aggregated explanation, is illustrated in Figure 1. Our approach leverages the network's inherent multi-head structure to identify and weight only the specific grid cells responsible for predictions. These sparse activations are then used to reconstruct exact, pixel-level receptive fields, which are ultimately aggregated using confidence scores to form the final saliency map. For development and experimentation, we rely on YOLOv11 [15], a widely used one-stage object detector.

3.1. Operating Regime and Target Selection

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input image. We operate strictly in the *preprocessing phase* of YOLOv11, before confidence thresholding and Non-Maximum Suppression (NMS). At this stage, the detector outputs fully differentiable raw predictions $\mathcal{D} = \{d_i\}_{i=1}^{N}$, where each $d_i = (x_i, y_i, w_i, h_i, c_i^{(1)}, \ldots, c_i^{(C)})$. A single forward pass yields activation tensors $\{\mathcal{A}^l\} = f(I)$. From $\mathcal{D}$, we select a target bounding box d_t to explain, with the target class defined inline as $c_t = \arg\max_k c_t^{(k)}$. Prior to NMS, YOLOv11 predicts multiple overlapping bounding boxes for the same object. To capture all valid internal evidence, we aggregate all boxes targeting the same object as d_t by filtering the raw predictions with a confidence threshold τ:

$$\mathcal{B}_t = \left\{ d_i \in \mathcal{D} \mid \mathrm{IoU}(d_t, d_i) > 0.45 \wedge \arg\max_k c_i^{(k)} = c_t \wedge c_i^{(c_t)} \geq \tau \right\}. \tag{1}$$

which captures all confident detections targeting the same object directly from the raw, pre-NMS outputs.
The $IoU > 0.45$ threshold follows the commonly used Non-Maximum Suppression (NMS) configuration in YOLO-based detectors, where highly overlapping boxes of the same class are assumed to correspond to the same object. By adopting this standard configuration, our aggregation mechanism mirrors the detector's internal logic for grouping multi-box predictions. Consequently, if a different NMS threshold is configured during inference, our method's threshold should simply be adjusted to match it. For each detection $d_i \in \mathcal{B}_t$, its index identifies the corresponding head, from which we extract the associated last convolutional layer $\mathcal{A}^{l_i}$ as the source for gradient computation. We define the scalar explanation target as the target class confidence $s_i = c_i^{(c_t)}$ and compute its gradient w.r.t the selected activations, $\mathcal{G}^{l_i} = \frac{\partial s_i}{\partial \mathcal{A}^{l_i}}$. Following HiResCAM [5], we compute the element-wise product and aggregate across K channels to obtain the weighted activation map (2), which is is highly sparse; only the active grid cells responsible for the detection are retained, significantly reducing computational cost.

$$\mathcal{H}^{l_i}(x, y) = \sum_{k=1}^{K} \left(\mathcal{A}_k^{l_i}(x, y) \cdot \mathcal{G}_k^{l_i}(x, y) \right). \tag{2}$$

3.2. Receptive Field Reconstruction and Saliency Aggregation

For each active grid cell (x, y) in $\mathcal{H}^{l_i}$, we reconstruct its receptive field via the input gradient $\mathcal{R}_{x,y} = \frac{\partial \mathcal{A}^{l_i}(x,y)}{\partial I}$. A smoothed pixel-level map is obtained by taking the L_2 norm across color channels and applying a Gaussian filter $\mathcal{G}_\sigma$:

$$\tilde{R}_{x,y} = \mathcal{G}_\sigma * \left\| \mathcal{R}_{x,y}(u, v, :) \right\|_2 . \tag{3}$$

We apply min-max normalization to project all receptive fields onto the same scale, ensuring they provide only spatial support while HiResCAM dictates their relative importance:

$$R_{x,y} = \frac{\tilde{R}_{x,y} - \min(\tilde{R}_{x,y})}{\max(\tilde{R}_{x,y}) - \min(\tilde{R}_{x,y})}, \quad R_{x,y} \in [0, 1]. \tag{4}$$

To aggregate the explanations across all bounding boxes in $\mathcal{B}_t$, we normalize their raw confidence scores s_i into a probability distribution using a softmax function, $\tilde{s}_i = \exp(s_i) / \sum_{d_j \in \mathcal{B}_t} \exp(s_j)$. The ultimate explanation S is then

constructed as the softmax-weighted sum of the individual detections:

$$S = \sum_{d_i \in \mathcal{B}_t} \left(\tilde{s}_i \sum_{(x,y)\in\Omega_i} \mathcal{H}^{l_i}(x,y) \cdot R_{x,y} \right). \tag{5}$$

This aggregation aligns with the detector's multi-box internal reasoning, yielding a faithful, object-centric explanation.

4. Experiments and Discussion

In this section, we present qualitative and quantitative evaluations of **RF-BoxCAM** and baseline techniques and discuss their strengths and limitations.

4.1. Experimental Setup

We evaluated the proposed method using the EAGLE dataset [16], a large-scale vehicle detection benchmark consisting of 8,280 high-resolution aerial images of 936×1404 pixels captured from real-world urban and highway scenarios. The dataset contains a variety of vehicle numbers and densities in each image, ranging from a few vehicles to around 2,500 vehicles, with two classes: large and small. Since the vehicles usually occupy only a small number of pixels, the EAGLE dataset is particularly suitable for studying small-object detection. Moreover, EAGLE is well suited for evaluating XAI techniques with respect to their faithfulness and spatial precision because it reflects realistic operational conditions with cluttered backgrounds, repetitive structures, and strong contextual dependencies.
For object detection, we considered YOLOv11 [15], a widely used, one-stage, deep learning–based method. Among its variants, we selected the large version, which performed better than the smaller versions on the EAGLE dataset in our preliminary model assessment. Since YOLOv11 accepts input images of 640×640 pixels, we cropped the original aerial images to this size with 10% overlap. We trained the model for 100 epochs on the EAGLE training set with a batch size of 32 and a learning rate of 0.01 for one class (merged the two classes). The trained model achieved a mean average precision mAP50 of 81% and mAP50-90 of 47% on the test set.
For qualitative and quantitative comparisons, we evaluate both white and black-box XAI methods. As white-box baselines, we consider Grad-CAM and HiResCAM applied to the class confidence score of each detection. As black-box baselines, we employ D-CLOSE and D-RISE, which generate explanations by measuring the effect of input perturbations on detector outputs.

4.2. Qualitative Evaluation

Figure 2 visualizes the saliency maps generated by different techniques for vehicle detections across various scenarios. To comprehensively evaluate these explanations, we divide the analysis into highly confident true positives (first five rows, confidence > 0.8) and low-confidence, ambiguous detections (last two rows, confidence < 0.45).

- **Highly Confident Detections:** For confident predictions, Grad-CAM typically scatters focus spatially over the entire scene, struggling to isolate small objects from high-frequency, repetitive background structures. HiResCAM improves spatial concentration but focuses too intensely on the center of the bounding box. This overlooks important peripheral evidence, an artifact of naively upsampling the activation space without considering the actual receptive field geometry of the multi-scale detector. Among the black-box methods, D-RISE generates object-aware maps but still exhibits noticeable background noise. D-CLOSE produces more spatially coherent areas that align with the expectation of a contiguous object. However, because these methods rely on random input perturbations, they provide only an approximate estimation of importance, typically resulting in a single large, contiguous blob of activation. In contrast, **RF-BoxCAM** not only precisely localizes the target but distinctly highlights multiple, spatially separated features (e.g., the windshield, rooftop, and cast shadows) that jointly drive the detection decision. By calculating gradients throughout the network and mapping them

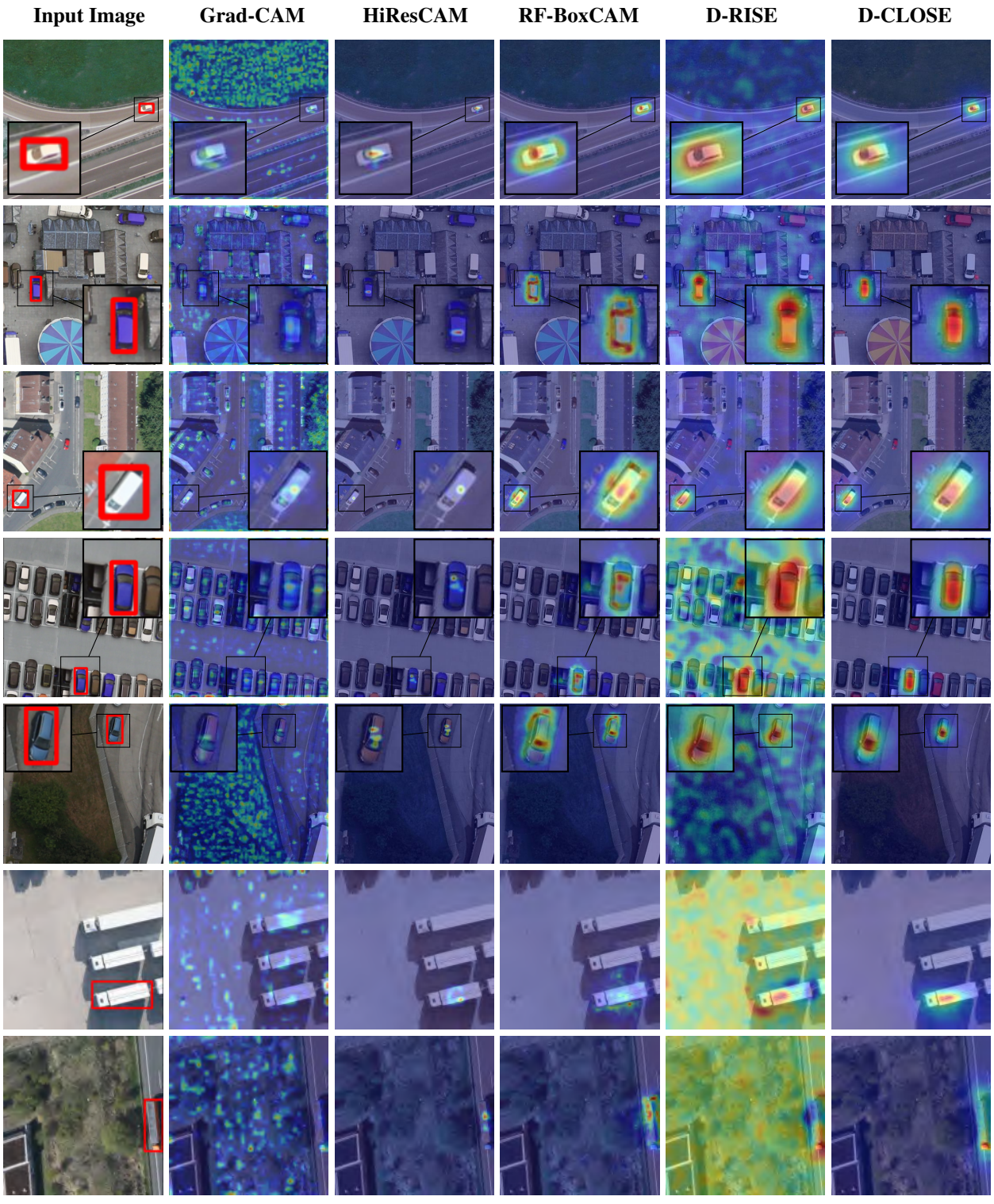

Fig. 2: Visualization of saliency maps produced by different techniques, including our **RF-BoxCAM**, for various sample detections from the EAGLE dataset. The target region in some examples is magnified for better visibility. All saliency maps are normalized to the range $[0, 1]$ and visualized using the *jet* colormap, where red indicates regions that are highly important for the detection decision and blue indicates regions that are not. The first five samples are detections with high confidence scores (> 0.8), while the last two samples are detections with low confidence scores (< 0.45).

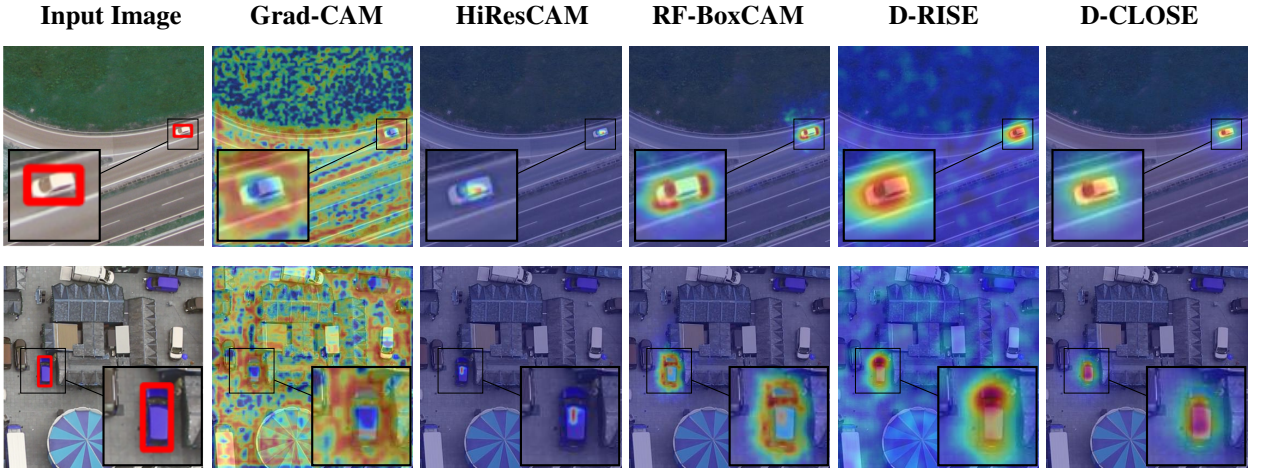

Fig. 3: Visualization of saliency maps produced by different techniques, including our **RF-BoxCAM**, for various sample detections from the EAGLE dataset when using YOLOv8.

to actual input-level receptive fields rather than perturbation-based approximations, RF-BoxCAM accurately captures how the model relies on distinct, non-contiguous areas to formulate its final prediction, providing a structural understanding that other techniques miss.

- **Low-Confidence Detections:** In the final two rows of Figure 2, Grad-CAM and HiResCAM produce scattered or misaligned saliency maps that fail to isolate the object. Similarly, black-box techniques provide coarse approximations that highlight the general vicinity but lack spatial precision. **RF-BoxCAM** isolates the vehicle while also identifying specific environmental features, such as road markings and rectangular shadows, that contribute to the model's lower confidence in these scenarios.

The main advantage of our **RF-BoxCAM** over HiResCAM lies in its ability to address the internal structure of modern multi-scale detectors on small objects. Instead of projecting the activation space directly onto the image space, **RF-BoxCAM** calculates gradients throughout the entire network and aggregates relevant information from various receptive fields across the input. We also compare **RF-BoxCAM** to baseline methods when using YOLOv8 to assess its stability across different YOLO architecture. Results are depicted in Figure 3, showing that **RF-BoxCAM**, HiResCAM, D-CLOSE and D-RISE produce highly consistent saliency maps across architectures, indicating stability within the YOLO family. In contrast, Grad-CAM exhibits more noticeable variations in the resulting explanations due to its reliance on coarse feature-map-level gradients, making it more sensitive to architectural differences.
For the false positive detections in Figure 4, Grad-CAM and HiResCAM produce scattered activations that fail to identify a specific cause for the error. Black-box methods generate broad, low-resolution regions that overlap with background textures. **RF-BoxCAM** highlights specific geometric structures in the environment, such as rectangular road features or building edges, that mimic the appearance of vehicles to the detector's internal layers, providing a clear visual explanation for the model's misclassification.
While most XAI technique evaluations rely only on feature importance for detection confidence, assessing the impact of these techniques on model decisions regarding the width and height of bounding boxes could also be important. Figure 5 shows the saliency maps generated by our method using the predicted width, height, and confidence score of the bounding box. As can be seen, the most important features are concentrated at the junctions of the car and trailer for all three scores. However, for box height, the model also focuses on the front and back of the trailer, where the top right and bottom left corners of the box are located.

4.3. Quantitative Evaluation

In the XAI domain, saliency maps are quantitatively evaluated for faithfulness using deletion and insertion metrics [23], which assess the causal importance of salient regions. In the deletion metric, image pixels are ranked by

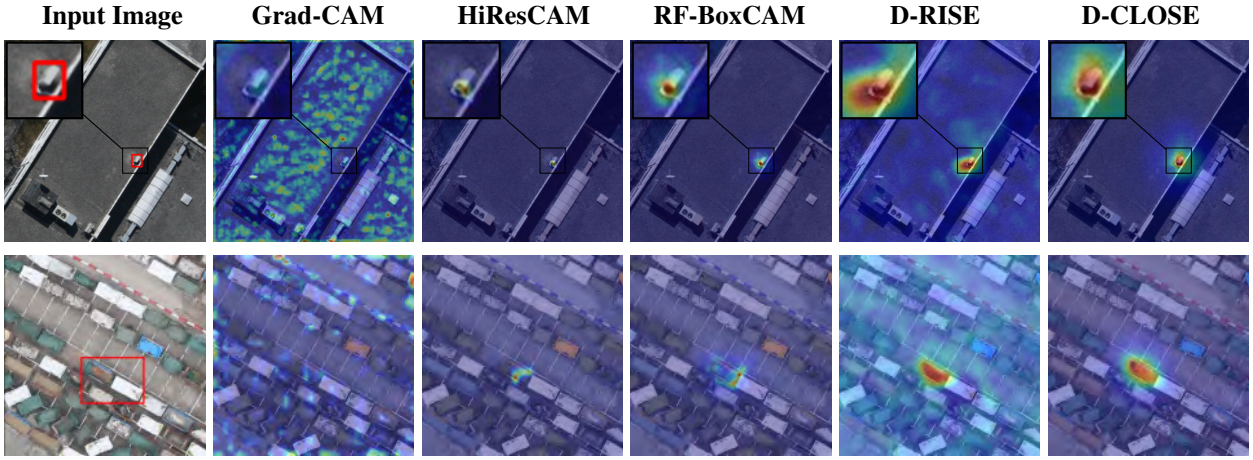

Fig. 4: Visualization of saliency maps for false positive detections from the EAGLE test set. The saliency maps are normalized to [0, 1] using the *jet* colormap (red indicates high importance, blue indicates low). The target region for the first sample is magnified for clarity.

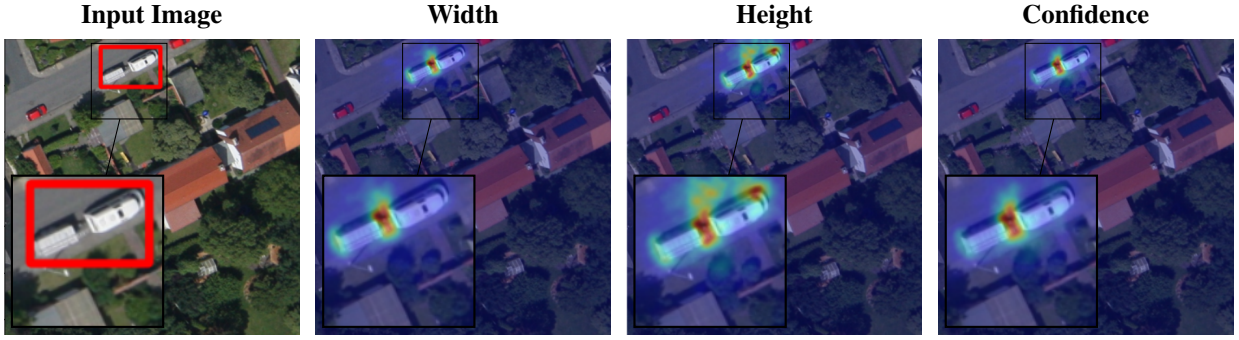

Fig. 5: Saliency maps produced by **RF-BoxCAM** when considering the width, height, and confidence information of the detection bounding box.

their saliency values. Starting with the most salient pixels, pixels are progressively removed from the input image. The detection confidence value is then measured, resulting in a confidence versus saliency curve. A faithful saliency map is indicated by a rapid and monotonous decrease in the curve and a small Area Under the Curve (AUC). The insertion metric also ranks pixels based on their saliency value. It starts from a blurred baseline image and progressively inserts the most salient pixels from the most salient one. The detection confidence score is measured. In this case, a faithful saliency map results in a steep increase in confidence value, resulting in a high AUC. The deletion and insertion metrics are usually combined by taking the difference between their AUCs, yielding a single scalar score that reflects the necessity and sufficiency of the highlighted regions.

For our evaluation, we calculated deletions, insertions, and their combinations for 3000 samples from the EAGLE test set. Additionally, we compared the time required to produce one saliency map using different techniques to evaluate their computational efficiency. The results are presented in Table 1. To analyze the performance of the XAI technique with respect to detection confidence, we divided the samples into two groups: high-confidence detections (> 0.80) and low-confidence detections (< 0.45). We assumed that these samples would be more useful for assessing the performance of the techniques.

Table 1 demonstrates that black-box techniques (D-Rise and D-Close) generate saliency maps that are significantly more accurate than those produced by the state-of-the-art white-box techniques (Grad-CAM and HiResCAM), for both high- and low-confidence detections (e.g., more than 83% versus 52.77% and 61.08% in the combined metric). However, black-box techniques require much longer computation times (more than 120 s) versus ~3 s for white-box ones. Table 1 shows also that the **RF-BoxCAM** saliency maps are as faithful as the other white-box techniques (the

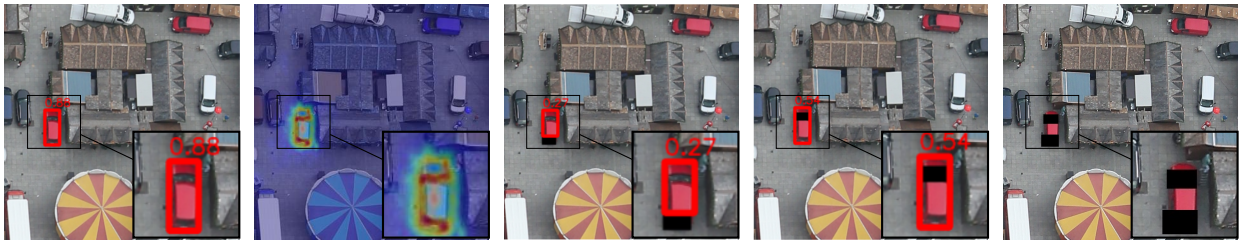

Fig. 6: Evaluating the real influence of the salient features provided by the proposed **RF-BoxCAM** method through manipulation of important features. From left to right: Original image with an overlaid detection bounding box and confidence score; saliency map by **RF-BoxCAM**; and three detection results with confidence scores after important features were manipulated by masking them with black patches.

second-best performing method in most cases), while maintaining a small computation time (6 s). **RF-BoxCAM** performs competitively with D-CLOSE but is 20× faster.

Metric	Grad-CAM	HiresCAM	**RF-BoxCAM**	D-RISE	D-CLOSE
	detection confidence > 0.80				
insertion [%] ↑	64.14	68.61	<u>84.31</u>	83.97	**84.76**
deletion [%] ↓	11.37	7.53	0.69	<u>0.23</u>	**0.19**
insertion–deletion [%] ↑	52.77	61.08	83.62	<u>83.74</u>	**84.57**
	detection confidence < 0.45				
insertion [%] ↑	13.02	20.80	<u>31.24</u>	25.07	**31.56**
deletion [%] ↓	0.81	0.16	<u>0.05</u>	0.64	**0.03**
insertion–deletion [%] ↑	12.22	20.64	<u>31.19</u>	24.43	**31.54**
runtime [s] ↓	**3**	<u>4</u>	6	180	120

Table 1: Quantitative evaluation of the faithfulness of saliency maps based on deletions, insertions, and their combinations on the test set of the EAGLE dataset. The last row shows the computation time for producing one saliency map. The best (resp. second best) performances are in bold (resp. underlined)

4.4. Empirical Study

To rigorously evaluate the accuracy of the explanation provided by the proposed **RF-BoxCAM** method, we manipulate different salient features and demonstrate the detection results in Figure 6. As can be seen in the figure, the saliency map shows two salient areas on the windshield and rear shield of the car. Masking the important features in the rear by a black patch changes the bounding box shape and decreases the confidence score from the baseline of 0.88 to 0.27. Masking the windshield decreases the confidence score to 0.54, but the bounding box shape remains the same. This is most likely due to other impactful features on the front of the vehicle, which are partially visible in the saliency map. When both areas are masked, the confidence score decreases drastically, and the model cannot detect the car.

5. Conclusions and Future work

This paper introduced **RF-BoxCAM**, a receptive-field guided explanation framework designed for object detection in aerial imagery, where objects are small and scenes are highly cluttered. By explicitly reconstructing the receptive field of each detection, the proposed approach generates object-centric explanations that remain spatially consistent with the detector's decision process.

Experimental results demonstrate that **RF-BoxCAM** successfully balances faithfulness and computational efficiency. The method achieves a level of explanation faithfulness comparable to state-of-the-art black-box approaches such as D-CLOSE, while providing a significant computational advantage, operating up to 20× faster. This efficiency makes the method highly suitable for large-scale offline analysis and efficient diagnostic auditing. Beyond efficiency, **RF-BoxCAM** enables diagnostic decoupling between classification and localization processes. By analyzing these components separately, the method helps reveal failure modes such as feature entanglement in dense urban scenes, where a detector may correctly identify an object class but inaccurately regress its bounding box. This diagnostic capability is particularly valuable for improving detectors deployed in safety-critical aerial monitoring and autonomous

systems. Despite these advantages, a primary limitation of **RF-BoxCAM** is its reliance on precise theoretical receptive field estimation, which inherently assumes the fixed, deterministic spatial structure of Convolutional Neural Networks (CNNs). Consequently, directly applying this framework to architectures governed by dynamic, input-dependent receptive fields such as Vision Transformers (ViTs), presents a challenge. In ViTs, the effective receptive field is continuously modulated by self-attention mechanisms rather than static convolutional layers. An important direction for future research is to extend this framework by reinterpreting dynamic attention maps and token interactions as adaptable receptive fields. This adaptation would allow our high-resolution explanation strategy to be seamlessly integrated with state-of-the-art transformer-based object detectors.

Acknowledgements

We thank the German Academic Exchange Service (DAAD) for supporting this research.

References

[1] Adadi, Amina, and Mohammed Berrada. (2018) "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." *IEEE access* **6**: 52138–52160.

[2] Andres, A., A. Martinez-Seras, I. Laña, and J. Del Ser. (2024) "On the black-box explainability of object detection models for safe and trustworthy industrial applications." *Results in Engineering* **24**: 103498.

[3] Simonyan, K., A. Vedaldi, and A. Zisserman. (2013) "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034*.

[4] Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. (2017) "Grad-CAM: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*: 618–626.

[5] Draelos, R. L., and L. Carin. (2020) "Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks." *arXiv preprint arXiv:2011.08891*.

[6] Luo, W., Y. Li, R. Urtasun, and R. Zemel. (2016) "Understanding the effective receptive field in deep convolutional neural networks." *Advances in Neural Information Processing Systems (NIPS)*: 4898–4906.

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

[8] Lundberg, S. M., and S.-I. Lee. (2017) "A unified approach to interpreting model predictions." *Advances in NeurIPS*: 4765–4774.

[9] Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. (2018) "A survey of methods for explaining black box models." *ACM Computing Surveys (CSUR)* **51** (5): 1–42.

[10] Petsiuk, V., A. Das, and K. Saenko. (2018) "RISE: Randomized input sampling for explanation of black-box models." *Proceedings of BMVC*.

[11] Petsiuk, V., R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, and K. Saenko. (2021) "Black-box explanation of object detectors via saliency maps." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*: 11443–11452.

[12] Zablocki, M., H. Ben-Younes, P. Pérez, and M. Cord. (2022) "Explainability of vision-based autonomous driving systems: Review and challenges." *International Journal of Computer Vision* **130**: 2425–2452.

[13] Lin, T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. (2017) "Feature pyramid networks for object detection." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 2117–2125.

[14] Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. (2016) "You only look once: Unified, real-time object detection." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 779–788.

[15] Ultralytics. (2024) "YOLOv11 Documentation." `https://docs.ultralytics.com/`.

[16] Azimi, S. M., R. Bahmanyar, C. Henry, and F. Kurz. (2021) "EAGLE: Large-scale vehicle detection dataset in real-world scenarios using aerial imagery." *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*: 6920–6927.

[17] Neubeck, A., and L. Van Gool. (2006) "Efficient non-maximum suppression." *Proceedings of the 18th International Conference on Pattern Recognition (ICPR)* **3**: 850–855.

[18] Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. (2020) "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* **58**: 82–115.

[19] Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. (2016) "Learning deep features for discriminative localization." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 2921–2929.

[20] Paszke, A., et al. (2019) "PyTorch: An imperative style, high-performance deep learning library." *Advances in Neural Information Processing Systems (NeurIPS)*: 8024–8035.

[21] Truong, V. B., T. T. H. Nguyen, V. T. K. Nguyen, Q. K. Nguyen, and Q. H. Cao. (2024) "Towards better explanations for object detection." *Proceedings of the 15th Asian Conference on Machine Learning (ACML)*: 1385–1400.

[22] Jung, H., and Y. Oh. (2021) "Towards better explanations of class activation mapping." *Proceedings of the IEEE/CVF ICCV*: 1879–1887.

[23] Nguyen, L. P. T., H. T. T. Nguyen, and H. Cao. (2025) "ODExAI: A comprehensive object detection explainable AI evaluation." *Proceedings of the 48th German Conference on Artificial Intelligence (KI)*.