Hybrid CNN-Based Border Pixel Extraction with Attention-Enhanced Feature Fusion and Explainable AI for Real-Time Traffic Object Detection

M. Koteswara Rao
Department of Information Technology, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana, India
mailtomkrao@gmail.com

D. Manju
Department of CSE (CyS, DS) and AI&DS, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
nuthana525@gmail.com (corresponding author)

K. Kishore Kumar
Department of Electronics and Communication Engineering, Faculty of Science and Technology (IcfaiTech), ICFAI Foundation for Higher Education, Hyderabad, India
kishorekamarajugadda@gmail.com

Rajesh Kumar Verma
Department of CSE, CMR Engineering College, Hyderabad, India
rajeshverma.hyd10@gmail.com

Padmini Debbarma
ICFAI, Hyderabad, India
debbarmapadmini.scholar24@ifheindia.org

Boda Sindhuja
Department of Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, India
sindhuboda777@gmail.com

Received: 14 December 2025 | Revised: 18 January 2026 | Accepted: 4 February 2026

 DOI: https://doi.org/10.48084/etasr.16948

ABSTRACT

This study presents a novel object recognition model, called Object Border Pixel Extraction with Precise Shape Detection using CNN (OBPE-PSD-CNN), to improve fine-grained object recognition in complicated traffic scenarios. In contrast to traditional deep learning-based object detection techniques that mainly use bounding boxes or coarse segmentation masks, the suggested one incorporates border pixel extraction and skeletal-contour-based shape modeling into the detection pipeline and allows the object to be represented precisely in terms of structure and geometry. In addition to enhancing discriminative ability in groups of visually similar objects, a hybrid deep feature fusion architecture based on ResNet-50 and AlexNet is proposed, and then an attention-enhanced classifier based on a Convolutional Block Attention Module (CBAM) is used to refine features adaptively, spatially, and channel-wise. Unlike the current literature where explainability is seen as a post-hoc analysis, the suggested framework entails Explainable AI (XAI) mechanisms in the architecture, offering visual pieces of evidence on model decisions in real-time. The

suggested OBPE-PSD-CNN is an end-to-end solution to object detection, accurate shape extraction, segmentation, classification, and temporal tracking. Large-scale testing on benchmark traffic data such as CityFlow and UA-DETRAC demonstrates that the suggested algorithm is more effective than the state-of-the-art strategies, especially in dense, heterogeneous, and visually demanding traffic cases.

Keywords-object detection; object border; object shape; human-computer interface; deep learning; convolutional neural networks; traffic monitoring

I. INTRODUCTION

Intelligent transportation systems, autonomous driving, and smart surveillance have led to the rise of applications of Computer Vision (CV) and Artificial Intelligence (AI), which in turn require robust object perception algorithms that can work in real-time. Object detection, segmentation, and tracking are essential to make the correct decision in complicated traffic conditions with occlusions, alterations of illumination, and high-density multi-camera views. To demonstrate the novelty and contribution of the proposed OBPE-PSD-CNN framework, its performance is compared to state-of-the-art approaches tested on common benchmark datasets. CityFlow and UA-DETRAC are popular in the study of traffic surveillance because they allow equitable, repeatable, and significant comparisons of the results between methods.

Although YOLO-based methods demonstrate good real-time performance, they mainly work using bounding-box representations and are not modeled explicitly to show object boundaries. Attention-based fusion techniques are better at discriminating features, but without structural shape modeling. In contrast, OBPE-PSD-CNN combines explicit pixel border extraction and skeletal/contour form modeling in a single framework. The proposed approach, which combines hybrid feature fusion, attention mechanisms, and Explainable AI (XAI) integration, proves effective and practically relevant to real-time traffic surveillance applications. The key contributions of this work are as follows:

- A novel CNN-based architecture for border pixel extraction and precise shape-based object representation.
- A hybrid feature fusion strategy that combines ResNet-50 and AlexNet for enhanced deep feature richness and robustness.
- A CBAM-based attention-enhanced deep classifier for improved feature discrimination.
- An end-to-end pipeline that incorporates detection, segmentation, classification, and temporal tracking.
- XAI techniques to ensure transparency and interpretability in decision-making.
- Comprehensive validation using large-scale real-world traffic datasets under varying environmental conditions.

Conventional end-to-end CNN edge detector designs have shown that a more explicit function of learning object boundaries allows more accurate instance-level detection by directly learning fine-grained border pixels of visual information [1]. Based on this concept, CNNs have been combined with task-specific edge detection, demonstrating that explicit boundary learning greatly improves segmentation and object delineation accuracy [2].

To further enhance feature representation, attention-enhanced multiscale feature fusion networks have also been suggested to effectively combine spatial and contextual information and achieve better segmentation performance in complex visual scenes [3]. Similarly, attention-based multiscale CNN feature fusion has been shown to enhance recognition strength in difficult visual scenarios such as occlusions and changes in illumination [4]. Regarding the topic of real-time traffic perception, systematic reviews of YOLO-based traffic sign detection algorithms reveal that there is a strong trade-off between detection accuracy and real-time capabilities, which makes the YOLO architecture a predictable baseline for time-sensitive applications [5]. Recent, more efficient YOLO versions, such as YOLOv8-CE, can be faster and more accurate in detecting objects in autonomous driving environments whilst being able to respect real-time prerequisites [6].

To resolve transparency issues in deep object detection systems, the discussion of black-box explainability approaches focuses on the role of XAI techniques in enhancing trust, safety, and interpretability in industrial and surveillance systems [7]. To effectively capture objects in a dynamic free traffic environment, deep stochastic configuration networks have been suggested to ensure that real-time detection efficiency is achieved with high accuracy at low computational costs [8]. In addition, attention-enhanced feature fusion strategies can be used to enhance discriminative feature learning using adaptive emphasis of informative spatial and channel-wise representations [9]. Adaptive deep learning strategies combined with DeepSORT in video surveillance have demonstrated higher accuracy in multi-object tracking with the promotion of temporal association [10]. Moreover, lightweight detection models, based on YOLOv7 with MobileNetv3, have been suggested to secure a trade-off between detection precision and computational efficiency, being applicable to real-time object detection tasks with scarce resources [11].

II. PROPOSED METHOD

The proposed Object Border Pixel Extraction with Accurate Shape Detection using Convolutional Neural Network (OBPE-PSD-CNN) is an integrated deep learning model that can be used to detect fine-grained objects in complex traffic conditions. Compared to existing methods, OBPE-PSD-CNN incorporates explicit border pixel learning and skeletal contour learning into the detection process. Its general architecture comprises the following sequential modules: (i) Input Preprocessing, (ii) Feature Extraction (ResNet-50+AlexNet), (iii) Hybrid Feature Fusion, (iv) Border Pixel Extraction Module, (v) Skeletal/Contour Shape Modeling, (vi) CBAM

Attention Refinement, (vii) Classification and Localization Head, and (viii) XAI Inference Module.

Figure 1 presents the framework of the proposed method.

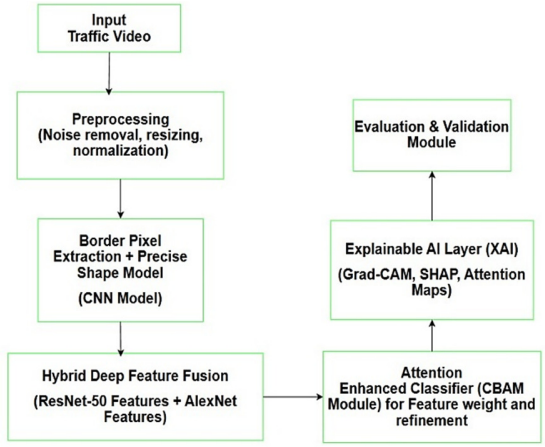

Fig. 1. Workflow of object detection and shape recognition process.

A. Input Preprocessing

The CityFlow [12] and UA-DETRAC [13] datasets were preprocessed to provide uniformity, stability, and enhanced generalization before the training of the OBPE-PSD-CNN model. Considering a given input RGB traffic image $I \in \mathbb{R}^{H\times W\times 3}$, the following preprocessing steps were applied:

- Resizing to 224×224
- Pixel normalization:

$$I_{norm} = I - \mu/\sigma \quad (1)$$

where μ and σ are the dataset's mean and standard deviation.

B. Feature Extraction

To enhance feature diversity, a hybrid backbone architecture is employed.

1) ResNet-50

A pretrained ResNet-50 extracts deep semantic features. Let:

$$F_r = ResNet50\ (I_{norm}) \quad (2)$$

Feature maps are extracted from the final convolutional block before global pooling. The output dimension is :

$$F_r \in R^{7\times 7\times 2048} \quad (3)$$

ResNet contributes strong hierarchical and residual feature learning.

2) AlexNet Branch

In parallel, AlexNet extracts complementary spatial texture features:

$$F_a = AlexNet(I_{norm}) \quad (4)$$

with its output dimension being:

$$F_a \in \mathbb{R}^{6\times 6\times 256} \quad (5)$$

AlexNet captures fine-grained edge and texture information useful for border extraction.

Since both datasets are initially offered as video sequences, the videos were initially transformed into single frames with a sampling rate of 10-15 Frames Per Second (FPS), which minimizes redundancy in time, yet maintains diversity in the scenes. The extracted frames were resized to 224×224 pixels to fit the input size of the backbone networks. To stabilize gradient updates in training, pixel intensities were normalized using the dataset mean and standard deviation. To make the model robust in different traffic conditions, data augmentation was used only in the training phase, with random horizontal flipping (probability 0.5), random brightness and contrast modifications, random scaling (between -10 and +10%), random cropping, and low-variance Gaussian noise insertion. Lastly, the data was split into 80% training, 10% validation, and 10% testing subsets to perform systematic hyperparameter-tuning and unbiased performance evaluation.

3) Hybrid Feature Fusion Module

To combine semantic depth and spatial detail, feature fusion is performed as follows:

1. Resize the AlexNet feature map to match the spatial dimension of ResNet:

$$F_a' = Resize(F_a)$$

2. Channel concatenation:

$$F_{fusion} = Concat(F_r, F_a')$$

3. 1×1 convolution for dimensionality reduction:

$$F_{hf} = Conv1\times 1(F_{fusion})$$

This produces: $F_{hf} \in \mathbb{R}^{7\times 7\times 1024}$. The 1×1 convolution acts as a learned weighted fusion mechanism.

Suppose that the input image is represented as $I \in \mathbb{R}^{224\times 224\times 3}$. Feature extraction is performed using a dual-backbone architecture to capture complementary semantic and spatial information. First, a ResNet-50 backbone extracts high-level semantic features from the final convolutional block (Conv5_x), producing a feature map $F_r \in \mathbb{R}^{7\times 7\times 2048}$. These features encode deep contextual representations essential for robust object recognition. In parallel, an AlexNet backbone extracts texture and edge-sensitive features from its final convolutional layer (Conv5), yielding $F_a' \in \mathbb{R}^{13\times 13\times 256}$. Since the spatial dimensions of the two feature maps differ, spatial alignment is performed by resizing the AlexNet feature map using bilinear interpolation, resulting in $F_a' \in \mathbb{R}^{7\times 7\times 256}$, ensuring compatibility with the ResNet feature map. The fusion strategy proceeds in three stages. First, channel-wise concatenation is applied, $F_{cat} = Concat(F_r, F_a')$, preserving complementary semantic and texture information. Second, a learnable dimensionality reduction step is performed using a 1×1 convolution with 1024 filters, followed by ReLU activation and Batch Normalization, producing a fused feature map. This 1×1 convolution enables adaptive weighted fusion of deep semantic and fine-grained spatial features. Third, regularization techniques, including Dropout (0.3) and Batch

Normalization, are applied to mitigate overfitting. Finally, the fused feature map is passed to the CBAM, $F_{att} = CBAM(F_{fusion})$, allowing adaptive refinement through sequential channel and spatial attention mechanisms, thereby enhancing discriminative feature representation before subsequent border extraction and classification stages.

C. Border Pixel Extraction Module

The border extraction branch explicitly learns object boundaries. Given the fused feature map F_{hf}, a convolutional edge detection head is applied:

$$B = \sigma(Conv3 \times 3(F_{hf}))$$

where $B \in \mathbb{R}^{7\times7\times1}$, and σ is sigmoid activation. Border supervision uses Binary Cross-Entropy loss:

$$L_{BCE} = -1/N\sum[ylog(\hat{y}) + (1-y)log(1-\hat{y})] \quad (6)$$

This branch forces the network to explicitly learn pixel-level boundary representations. The border extraction module is applied as a specialized convolutional branch. The network has three consecutive convoluted layers that are used to narrow down to boundary representations. The first convolutional layer consists of a 3×3 kernel and 512 filters, with same padding, and ReLU activation. The second convolutional layer uses a 3×3 filter size and 256 filters, after which ReLU activation and Batch Normalization are used to stabilize the learning process and enhance convergence. The last predictive layer is the border prediction, where a 1×1 convolution is implemented with a single filter, and sigmoid activation is applied to produce a pixel-wise-border probability map. The border extraction module is an exclusive convolutional branch that uses the fused feature map to process the border extraction module.

D. Skeletal/Contour Shape Modeling

To enhance geometric understanding, a skeletonized contour representation is generated from a predicted border map B, $S = Skeletonize(B)$, where morphological thinning is applied. Skeleton pixels encode the central structural axis. The shape descriptor vector is given by $Ds = GlobalAvgPool(S)$. This descriptor captures geometric information independent of bounding box approximation.

E. CBAM Attention Refinement

To refine fused features, a CBAM is integrated after feature fusion. Channel Attention is given by:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (7)$$

with refined features:

$$F_c = M_c(F) \otimes F \quad (8)$$

Spatial Attention:

$$M_s(F_c) = \sigma(Conv7 \times 7([AvgPool(F_c); MaxPool(F_c)])) \quad (9)$$

The final attention output is:

$$F_{att} = M_s(F_c) \otimes F_c \quad (10)$$

This enables adaptive emphasis on informative spatial and channel regions. Grad-CBAM is applied to the attention-refined convolutional feature maps, producing class-discriminative localization heatmaps. The procedure adds no more trainable parameters and reuses already available feature maps, such that the process is real-time applicable.

F. Classification and Localization Head

The refined features F_{att} are passed to a Fully Connected Layer (512 neurons), ReLU activation, Dropout (0.5), and a final Softmax classifier. Classification loss is given by:

$$L_{cls} = -\sum y_i \ log\ (\hat{y})$$

and the bounding box regression loss is given by:

$$L_{bbox} = Smooth_{L1}(b, \hat{y})$$

The proposed OBPE-PSD-CNN model takes RGB images as inputs. Since CityFlow and UA-DETRAC are video datasets, all video sequences were converted into single frames before training. A controlled frame rate was used to extract frames, which were treated as independent image samples. Following this preprocessing step, the two datasets were integrated into a uniform image format. Each frame was examined by the CNN separately. Tracking is implemented on top of frame-level detection and has no impact on the CNN input format.

G. Multi-Task Loss Function

The main task of the proposed OBPE-PSD-CNN model is defined as a multi-task loss function, which simultaneously optimizes classification, localization, and border extraction. The total loss is defined as:

$$L_{total} = \alpha L_{cls} + \beta L_{bbox} + \gamma L_{border} \quad (11)$$

where L_{cls} represents classification loss, L_{bbox} is the regression loss of the bounding box, and L_{border} is the border extraction loss. Empirical values were set for the weighting coefficients: α=1.0, β=1.0, γ=0.5. This ensures balanced detection and shape representation learning. γ was set to 0.5 to balance object recognition learning and accurate shape representation. Such a weighted formulation permits the network to have a high level of detection accuracy and also learn a correct boundary representation, hence structural awareness is enhanced without saturating the main detection goal.

H. Training Strategy

The Adam optimizer, which is renowned for its effective management of sparse gradients and adaptive learning rates, was used to train the model. Stable convergence was ensured by setting the initial learning rate at 0.0001, which provided a conservative step size. A batch size of 32 was used for training to balance gradient stability and computational efficiency. The model was trained for 100 epochs to give it enough iterations to learn intricate patterns from the data. Step decay was used to implement a learning rate scheduler that lowers the learning rate every 30 epochs to enhance convergence and avoid overfitting. An NVIDIA GPU was used to speed up all computations, reducing training time and making it possible to

handle big datasets effectively. Transfer learning was applied by initializing backbones with pretrained ImageNet weights.

The experience with the implementation of XAI as part of the system proves that model transparency and the ability to operate in real-time do not exclude each other. The proposed system can obtain interpretable predictions without causing latency or computation bottlenecks by using lightweight and gradient-based explanation methods that reuse the network representations of intermediate layers. This offers a balance, especially in traffic monitoring and safety applications, where both transparency in decision-making and timely response are mandatory.

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Benchmark Datasets

This study used benchmark datasets that were carefully chosen to represent a range of circumstances of interest in traffic jam reduction. These datasets comprise an assortment of external variables, including different twilight hours, weather conditions, and traffic levels, to ensure accuracy, diversity, and robustness.

1) Traffic Dataset A

CityFlow [12] is a large-scale real-world traffic surveillance dataset with video streams from multiple cameras at intersections and urban roads. The dataset includes traffic scenarios coupled with diverse traffic participants, such as vehicles, pedestrians, and traffic signals, annotated, and is widely used for benchmarking various detection, tracking, and congestion-control systems.

2) Traffic Dataset B

UA-DETRAC [13] consists of real videos of surveillance footage, recorded under various lighting and weather conditions, including heavy traffic and occlusion. There are annotations of various types of vehicles, including cars, trucks, vans, and buses. This dataset is ideal for testing the performance of an object detection and multi-object tracking system for traffic monitoring.

B. Training and Testing Data

Experimentation was based on the UA-DETRAC dataset [13], having a total of 5805 images, with 80% (4644 images) used to train the model and 20% (1161 images) to test it.

C. Evaluation Metrics

A set of typical evaluation metrics from the area of object detection was employed to evaluate the performance of the OBPE-PSD-CNN model. Among these measures were:

- Precision: The number of correctly identified items divided by the total number of recognized ones.
- Recall indicates how well a model identifies all relevant items, being the ratio of correctly identified objects to all objects in reality.
- F1-score is the harmonic average of Recall and Precision, measuring both the algorithm's precision and completeness.
- Mean Average Precision (mAP), which encompasses a precision and recall trade-off over many types of items. This provides a complete assessment of model accuracy.

D. Precision-Recall Curves

Figures 2-4 present Precision-Recall curves for each item category to demonstrate how the model performs. These curves help to clarify the model's capabilities for various traffic-related items by showing the trade-off between recall and accuracy.

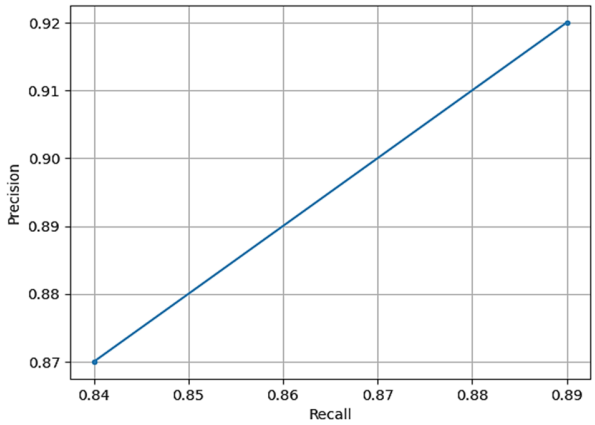

Fig. 2. Precision-Recall curve for vehicles.

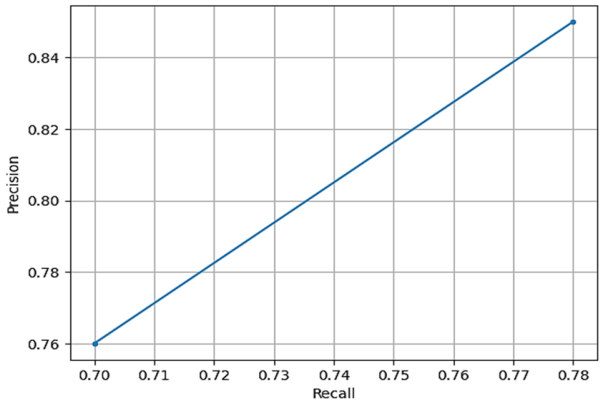

Fig. 3. Precision-Recall curve for pedestrians.

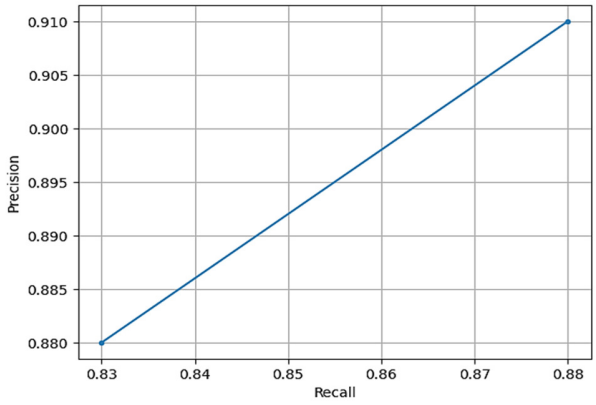

Fig. 4. Precision-Recall curve for traffic signs.

Figure 2 shows that the model has a strong level of precision in relation to recall for the classification of vehicles. The steady increase in recall values from 0.84 to 0.89 shows that the corresponding values of precision have also been steadily increasing from 0.88 to 0.92. Figure 3 shows that regarding pedestrians, precision levels up from 0.77 to 0.85, with recall varying from 0.70 to 0.78. Note that compared to vehicles, the performance of the model is slightly lower.

Table I summarizes the effectiveness of the proposed object recognition model on the two datasets. These results show that the proposed OBPE-PSD-CNN model is a potent tool for improving urban traffic management and safety while also excelling in real-time object recognition for traffic congestion reduction. The proposed model performed well on both the Traffic Dataset A, which includes normal scenarios of traffic flow, and the Traffic Dataset B, including more challenging situations. High F1-scores indicate that there is a balance between accuracy and recall, qualifying the proposed algorithm for real traffic monitoring and congestion management applications. It also gives evidence of the robustness of the model across different datasets for its potential to be implemented in various urban areas.

TABLE I. SUMMARY OF PERFORMANCE METRICS

Dataset	Object category	Precision	Recall	F1-score
Dataset A [12]	Vehicles	0.92	0.89	0.91
	Pedestrians	0.85	0.78	0.81
	Traffic Signs	0.91	0.88	0.89
Dataset B [13]	Vehicles	0.87	0.84	0.85
	Pedestrians	0.76	0.70	0.73
	Traffic Signs	0.88	0.83	0.85

The XAI module functions only during the inference stage and uses previously created feature maps and attention responses created by the trained network. Adding XAI does not have an impact on the convergence of a model or its inference speed since no other parameters or re-training are needed, and thus it is applicable to real-time deployment.

E. Implications for Traffic Congestion Control

The proposed OBPE-PSD-CNN model may lead to improved traffic control measures, since it provides real-time object detection and recognition, enhancing situational awareness of any traffic control system. This can lead to the development of better decision-making and response mechanisms in urban areas to reduce traffic congestion, improve road safety, and optimize mobility.

IV. CONCLUSION

The proposed OBPE-PSD-CNN model showed promise, achieving an accuracy of 0.92, a recall of 0.89, and an F1-score of 0.91 when applied in Traffic Dataset A, and accuracy, recall, and F1-score values of 0.85, 0.78, and 0.81, respectively, on the trickier Traffic Dataset B. Notably, the model demonstrated exceptional accuracy (0.91), recall (0.88), and an F1-score of 0.89 in the domain of traffic signs. These results demonstrate how well the OBPE-PSD-CNN model performs in accurately detecting and classifying objects in a variety of traffic settings, highlighting its potential to dramatically improve traffic congestion management techniques, opening up a viable field for additional investigation and implementation.

Further exploration of object skeleton extraction methods and the integration of advanced technologies like image enhancement and super-resolution will be essential. In addition, the potential of the OBPE-PSD-CNN model in various applications, such as autonomous vehicles or medical monitoring, opens doors to extend research in these areas. Even though the proposed framework shows great results on different benchmark datasets, future research should investigate training it on larger and more diverse datasets and further optimizing it to further improve performance. By providing a wider area of training data with more scenarios of traffic, categories of objects, and environmental factors, it is likely to increase the generalization capacity and strength of the model.

REFERENCES

[1] X. Zou, H. Liu, and Y. J. Lee, "End-to-End Instance Edge Detection." arXiv, 2022, https://doi.org/10.48550/ARXIV.2204.02898.

[2] L. C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform." arXiv, 2015, https://doi.org/10.48550/ARXIV.1511.03328.

[3] K. Dong *et al.*, "Attention-enhanced multiscale feature fusion network for pancreas and tumor segmentation," *Medical Physics*, vol. 51, no. 12, pp. 8999–9016, Dec. 2024, https://doi.org/10.1002/mp.17385.

[4] M. Xiao, B. Yang, S. Wang, Z. Zhang, X. Tang, and L. Kang, "A feature fusion enhanced multiscale CNN with attention mechanism for spot-welding surface appearance recognition," *Computers in Industry*, vol. 135, Feb. 2022, Art. no. 103583, https://doi.org/10.1016/j.compind.2021.103583.

[5] M. Flores-Calero *et al.*, "Traffic Sign Detection and Recognition Using YOLO Object Detection Algorithm: A Systematic Review," *Mathematics*, vol. 12, no. 2, Jan. 2024, https://doi.org/10.3390/math12020297.

[6] Y. Luo, Y. Ci, H. Zhang, and L. Wu, "A YOLOv8-CE-based real-time traffic sign detection and identification method for autonomous vehicles," *Digital Transportation and Safety*, vol. 3, no. 3, pp. 82–91, 2024, https://doi.org/10.48130/dts-0024-0009.

[7] A. Andres, A. Martinez-Seras, I. Laña, and J. Del Ser, "On the black-box explainability of object detection models for safe and trustworthy industrial applications," *Results in Engineering*, vol. 24, Dec. 2024, Art. no. 103498, https://doi.org/10.1016/j.rineng.2024.103498.

[8] Y. Wang, Y. Liu, R. Yi, and Y. Jiang, "Real-time traffic object detection algorithm with deep stochastic configuration networks," *Information Sciences*, vol. 700, May 2025, Art. no. 121848, https://doi.org/10.1016/j.ins.2024.121848.

[9] Y. Zheng, S. Liu, and L. Bruzzone, "An Attention-Enhanced Feature Fusion Network (AeF^2 N) for Hyperspectral Image Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023, https://doi.org/10.1109/LGRS.2023.3320193.

[10] M. K. Rao and P. M. A. Kumar, "Advanced Object Tracking in Video Surveillance Systems with Adaptive Deep SORT Enhancement," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 20871–20877, Apr. 2025, https://doi.org/10.48084/etasr.9529.

[11] L. Jia *et al.*, "MobileNet-CA-YOLO: An Improved YOLOv7 Based on the MobileNetV3 and Attention Mechanism for Rice Pests and Diseases Detection," *Agriculture*, vol. 13, no. 7, Jun. 2023, https://doi.org/10.3390/agriculture13071285.

[12] "2022 Data and Evaluation – AI City Challenge." https://www.aicitychallenge.org/2022-data-and-evaluation/.

[13] "UA-DETRAC_dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/dtrnngc/ua-detrac-dataset.