# Enhancing Vision-Based Perception in Autonomous Driving: YOLO11–DETR Integration with Selection Model

Ahmed M. Reda[1,2]*, Naser El sheimy[1], Adel Moussa[1,3]

[1]Dept. of Geomatics Engineering, University of Calgary, Calgary, AB, Canada – (ahmed.hegazy@ucalgary.ca, elsheimy@ucalgary.ca, amelsaye@ucalgary.ca)
[2]Dept. of Geomatics Engineering, Benha University, Benha, Egypt – (ahmed.hegazy@ucalgary.ca)
[3]Dept. of Electrical and Computer Engineering, Port-Said University, Port-Said, Egypt – (amelsaye@ucalgary.ca)

**KEYWORDS:** Autonomous Driving, Object Detection, Selection Model, Domain Shift, YOLO11, RT-DETR.

**ABSTRACT:**

Vision-based object detection is a key component of autonomous driving perception systems; however, models pretrained on large-scale generic datasets usually struggles when implemented in automotive environments due to domain shift. This research introduces a comprehensive evaluation and fusion of YOLO11 and RT-DETR for improving robustness in autonomous driving scenarios using KITTI dataset. Both models are pretrained on COCO dataset and evaluated under a zero-shot transfer setting to assess cross-domain generalization. The results show that RT-DETR-L and RT-DETR-XL experience significant performance degradation, dropping from 53.0 and 54.8 $mAP$ on COCO to 34.3 and 34.5 on KITTI, respectively. In contrast, YOLO11-Nano and YOLO11-L demonstrate better generalization, achieving 44.0 and 51.3 $mAP$ on KITTI compared to 40.9 and 55.0 on COCO. Controlled fine-tuning experiments (10 and 100 epochs) are conducted to analyze adaptation dynamics. RT-DETR-L improves to 61.7 and 79.1 $mAP$, while YOLO11-L reaches 64.9 and 76.2 after 10 and 100 epochs, respectively. To further evaluate robustness under challenging conditions, three degraded data subsets are generated. Building on the strengths of convolutional and transformer-based detectors, this work introduces an image-based selection model that selects the most suitable detector for each input image. Experimental results demonstrate substantial zero-shot degradation, strong recovery after fine-tuning, and consistent performance improvements under degraded conditions using the proposed selection strategy. Our method achieves gains of up to 4 $mAP$ points over the best standalone detector without incurring the computational overhead. The proposed framework provides a context-aware and computationally efficient perception enhancement strategy suitable for real-world autonomous driving systems.

## 1. INTRODUCTION

Autonomous driving has emerged as a transformative technology with significant potential to enhance road safety and transportation efficiency. However, for autonomous vehicles (AVs) to be viable in real-world deployment, they must operate reliably under diverse and challenging environmental conditions. These include nighttime driving, different conditions, and dynamic traffic scenarios which requiring rapid response to unexpected events. In addition, AV systems must maintain operational robustness even under partial sensor degradation or failure. Addressing these challenges is essential for the widespread adoption of autonomous driving technologies.

Modern AV systems rely heavily on perception modules to interpret the surrounding environment. Camera-based sensors offer a cost-effective solution with rich semantic information but are sensitive to illumination changes and lack direct depth estimation (Arnold et al., 2019). In contrast, Light Detection and Ranging (LiDAR) systems provide accurate geometric measurements but remain expensive and can also degrade under adverse weather conditions (Yoneda et al., 2019). Consequently, sensor fusion approaches that combine complementary sensing modalities have been widely explored to improve perception accuracy and robustness (Yao et al., 2024). Nevertheless, camera-based perception remains a core component due to its scalability and affordability.

Within vision-based perception, multiple tasks including object detection, tracking, lane detection, and semantic segmentation are integrated to achieve a comprehensive understanding of the driving environment. Among these, object detection plays a critical role, as it enables AVs to identify and localize key elements such as vehicles, pedestrians, traffic signs, and road infrastructure. Accurate object detection is fundamental for situational awareness, decision-making, and safe navigation, as well as for downstream tasks such as trajectory planning and behavior prediction (Gragnaniello et al., 2023).

Despite these advances, a key limitation remains, models trained on large-scale general-purpose datasets (e.g., COCO) often fail to generalize effectively to domain-specific environments such as autonomous driving scenarios. This performance degradation, commonly referred to as domain shift, arises from differences in viewpoint, object scale, scene structure, and environmental conditions. As a result, models that perform well on benchmark datasets may exhibit significant accuracy drops when deployed in real-world driving contexts.

To address this challenge, this study investigates the cross-domain generalization and adaptation behavior of two state-of-the-art object detection architectures: YOLO11 and RT-DETR. These models represent two distinct paradigms, convolutional and transformer-based detection, each with unique strengths in capturing local and global features. Furthermore, this work proposes an image-based selection model that chooses the most suitable detector for each input image, leveraging their complementary characteristics.

## 2. OBJECT DETECTION

The main objective of 2D object detection is detection and localization of objects in two-dimensional space through prediction of the bounding boxes related to these objects depending on sensory inputs in driving scenarios. This process is fundamental for autonomous systems to accurately understand

the surroundings and make safe and efficient navigation. 2D object detection can be mathematically expressed as a general formula which describe the relationship between both input sensor data and the 2D predicted bounding boxes:

$$\mathrm{B} = f_{det}(I_{sensor}) \quad (1)$$

In this context, $\mathrm{B} = \{B_1, \dots\dots\dots, B_N\}$ defines a set of detected objects within a scene, where $\boldsymbol{f_{det}}$ is the detection model and $\boldsymbol{I_{sensor}}$ refers to the sensory inputs. A key aspect of this detection process is defining how each object $\boldsymbol{B_i}$ is represented, as this impacts directly downstream tasks such as prediction and planning. Typically, objects are represented as a bounding box.

$$\mathrm{B} = [X_c, Y_c, w, h,\ \theta, class] \quad (2)$$

In this context, $(X_c, Y_c)$ define the center coordinates of the bounding box, while $w\ and\ h$ denote the object’s size which is width, and height, respectively. The optional parameter θ indicating the orientation. The term "class" defines the object’s category (cars, trucks, pedestrians, or cyclists, etc.).

Deep learning plays a critical role in perception systems tasks of AV which are detection and tracking objects, enable enhancement of environmental awareness and improvement of navigation safety. Most deep learning models are formed from deep convolutional neural networks (CNNs) where they analyse visual data like images or video frames for extracting relevant features, then these extracted features enable the system in detecting objects and estimating its location within the scene (Tardioli et al., 2023).

Traditional models of object detection usually struggle with occlusions, varying in lighting conditions, and environmental noise. In contrast, recent modern deep learning models, including transformer-based models and convolutional neural networks (CNNs), can effectively extract high-level semantic and geometric features, enabling the AV system to detect partially occluded objects and adapt to challenging driving conditions. Another major advantage of deep learning-driven detection in AVs is its capability of facilitating decision-making in real-time. Autonomous driving systems requesting fast and accurate perception which is essential in dynamic environments. Models like YOLO, Faster R-CNN, and transformer-based framework like TransFusion, achieve a balance between computational efficiency and detection accuracy. These models rely on various techniques like feature extraction, attention mechanisms, and multi-scale representations for improving capabilities of detection and tracking objects. Moreover, ongoing research areas are focused on optimizing these architectures to reduce computational overhead and enabling their deployment on systems on embedded hardware with limited resources (Bai et al., 2022).

## 3. STATE OF THE ART

Camera-based object detection has become a central research focus in autonomous driving due to advances in deep learning and computer vision. Despite the inherent limitations of image data, particularly the lack of explicit depth information, significant progress has been made in improving detection accuracy and robustness under complex driving conditions. Existing approaches can be broadly categorized based on their sensing configuration and data representation, including monocular, stereo, pseudo-LiDAR, multi-view, and Bird’s Eye View (BEV) methods.

Camera-based perception plays a critical role in tasks such as object detection, tracking, segmentation, and depth estimation. Its main advantages include low cost, scalability, and rich semantic information. However, the absence of direct depth measurements remains a key limitation, motivating the development of methods that infer or approximate 3D structure from image data.

### 3.1 Monocular Object Detection

Monocular approaches rely on a single camera to detect and localize objects. These methods are attractive due to their simplicity and cost-effectiveness but are inherently limited by the absence of depth information, which affects localization accuracy. Recent works have addressed this limitation using advanced representations and learning strategies. For example, MonoNeRD (Xu et al., 2023) integrates neural radiance fields with object detection, while MonoSKD (Wang and Zheng, 2023) employs self-knowledge distillation to enhance performance. More recent frameworks such as SSD-MonoDETR (He et al., 2024) demonstrate the growing effectiveness of transformer-based monocular detection.

### 3.2 Stereo Camera Object Detection

Stereo-based methods use multiple cameras to estimate depth through disparity between corresponding image points. These approaches provide a cost-effective alternative to LiDAR for depth estimation and can improve localization accuracy in both indoor and outdoor environments. Notable works include YOLOStereo3D (Liu et al., 2021), which integrates stereo matching into a YOLO framework, and LIGA-Stereo (Guo et al., 2021), which incorporates LiDAR geometry-aware features. Additionally, SIDE (Peng et al., 2022) introduces structure-aware depth estimation to enhance detection performance. Despite these advances, stereo methods remain sensitive to illumination changes and textureless regions.

### 3.3 Pseudo-LiDAR Object Detection

Pseudo-LiDAR approaches aim to bridge the gap between image-based and LiDAR-based detection by converting estimated depth maps into point cloud representations. These methods significantly improve detection performance compared to purely image-based approaches, particularly at longer ranges. For instance, Neighbor-Vote (Chu et al., 2021) improves point cloud consistency through local voting, while DD3D (Park et al., 2021) focuses on dense depth prediction. Pseudo-Stereo (Chen et al., 2022) further enhances performance by refining depth estimation and point cloud generation. However, their accuracy is still limited by errors in depth estimation, especially near object boundaries.

### 3.4 Multi-View Object Detection

Multi-view approaches utilize images captured from different viewpoints to improve scene understanding. These methods are computationally efficient and suitable for large-scale deployment but face challenges in accurate depth estimation and spatial consistency. ORA3D (Roh et al., 2023) introduces occlusion-aware reasoning to improve robustness, while PolarFormer (Jiang et al., 2023a) leverages polar-coordinate transformers for multi-camera perception. More recent methods, such as Vampire (Xu et al., 2024), focus on view-adaptive feature interaction to enhance detection accuracy across viewpoints.

### 3.5 Bird's-Eye View (BEV) Object Detection

BEV-based approaches transform image features into a top-down representation of the scene, enabling more accurate spatial reasoning for autonomous driving tasks. These methods are widely used in both camera-based and multi-modal perception systems. Recent developments emphasize transformer-based architectures that improve spatial-temporal modeling. BEVFormer v2 (Yang et al., 2023) introduces perspective supervision for improved convergence, while BEVDepth (Li et al., 2023) enhances depth estimation accuracy. SOGDet (Zhou et al., 2024) further improves robustness under challenging conditions by integrating semantic and occupancy information.

### 3.6 Summary and Research Gap

Although significant progress has been achieved across these approaches, most existing methods focus on improving detection performance through architectural design or sensor configuration. However, limited attention has been given to adaptive model selection strategies that leverage the strengths of different detection paradigms under varying environmental conditions. In particular, convolutional and transformer-based detectors exhibit complementary behaviors in terms of generalization, robustness, and adaptation. This observation motivates the need for a lightweight, context-aware framework that can exploit these complementary strengths without increasing computational complexity.

In this work, we address this gap by proposing an image-based selection model that adaptively chooses the most suitable detector for each input image, improving robustness under domain shift and degraded visual conditions.

## 4. DATA SET (KITTI)

The KITTI object detection benchmark is widely used for evaluating perception algorithms in autonomous driving scenarios. It consists of real-world data collected using a vehicle-mounted platform equipped with synchronized sensors, including RGB cameras and LiDAR. The dataset provides annotated images with object labels, 2D bounding boxes, and calibration parameters, enabling both image-based and multi-modal perception research (Geiger et al., 2013).

In this study, we focus on the camera-based object detection on KITTI. The dataset includes multiple object categories relevant to driving environments, such as cars, pedestrians, and cyclists, with annotations categorized based on object size, occlusion, and truncation. These characteristics make KITTI suitable for evaluating detection performance under realistic urban and semi-urban conditions.

Compared to the COCO dataset used for pretraining, KITTI exhibits significant domain differences, including a narrower field of view, consistent forward-facing camera geometry, different object scale distributions, and a limited set of traffic-specific classes. These differences introduce a notable domain shift, making KITTI an appropriate benchmark for evaluating cross-domain generalization.

To further assess robustness under challenging conditions, three degraded subsets of the KITTI validation data are generated. These subsets simulate adverse visual conditions through controlled perturbations, including illumination variation, brightness adjustment, noise injection, blur, and structural distortions. This setup enables a systematic evaluation of model robustness to realistic perception degradations beyond standard benchmark conditions.

## 5. METHODOLOGY

This main goal of this study is to analyse cross-domain generalization, adaptation behavior, robustness under visual degradation, and adaptive model selection. The overall framework consists of four main stages: (1) baseline evaluation of pretrained models, (2) fine-tuning on the target domain, (3) robustness assessment under synthetic visual degradations, and (4) development of image-based selection model for adaptive model arbitration.

### 5.1 Detection models

Two advanced object detection architectures were selected: YOLO11 and RT-DETR. The key motivation of choosing both models is their distinct architectural paradigms, that offer complementary strengths. YOLO11 is a convolutional one-stage detector optimized for real-time applications. Its architecture includes a backbone for feature extraction, a feature aggregation neck which allows multi-scale detection, and an anchor-free head for direct bounding box prediction as shown in Figure 1. Its convolutional networks inherently enable effective learning of local spatial features like textures. On the COCO benchmark, YOLO11-L achieves 55 $mAP_{(50-95)}$, while YOLO11-Nano achieves 40.9, demonstrating strong scalability across model sizes (Zhang et al., 2025).

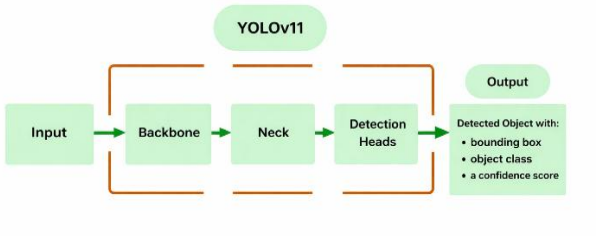


**Figure 1.** Architecture of the YOLO11 object detection model

In contrast, RT-DETR is a real-time adaptation of the detection transformer framework, which formulates object detection as a direct set prediction problem. Its architecture combines a convolutional backbone with a transformer encoder-decoder structure that uses learnable object queries and global self-attention mechanisms as shown in figure 2. Unlike traditional detectors, RT-DETR eliminates non-maximum suppression by employing Hungarian matching during training to enforce one-to-one prediction assignment. Transformer-based detectors excel at modeling global contextual relationships and long-range spatial dependencies. On COCO, RT-DETR-L achieves 53 $mAP_{(50-95)}$, and RT-DETR-XL reaches 54.8 (Zhao et al., 2023).

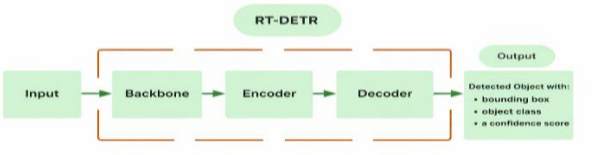


**Figure 2.** Architecture of the RT-DETR object detection model

### 5.2 Experiments

All detectors are initialized from COCO-pretrained weights provided by their respective frameworks, ensuring a common starting point in terms of training data and label taxonomy. The target domain is the KITTI object detection benchmark, which consists of forward-facing camera images captured from a vehicle platform in urban and semi-urban environments; KITTI differs from COCO in geometry, viewpoint, object scale distribution, and the predominance of traffic-related classes. Evaluation is performed using $mAP_{(50-95)}$, computed across standard IoU thresholds between 0.50 and 0.95, to align with COCO-style metrics while reflecting localization quality.

**5.2.1 Zero-shot evaluation on KITTI:** evaluates both models in a zero-shot transfer setting from COCO to the KITTI object detection dataset. This setup quantifies pure cross-domain transfer from general-purpose scenes to driving imagery and isolates the domain shift effect, as no retraining or adaptation is performed.

**5.2.2 Fine-tuning on standard KITTI:** RT-DETR-L and YOLO11-L are fine-tuned using clean KITTI images under two training regimes: short adaptation (10 epochs) and extended adaptation (100 epochs). This allows analysis of early convergence behavior and long-term adaptation capacity. After each regime, the models are evaluated on the clean KITTI validation set.

**5.2.3 Robustness evaluation on generated KITTI samples:** To assess the robustness under perception challenges, we generated three main degraded subsets of data (Sample 1, Sample 2, and Sample 3) by adding controlled visual perturbations such as illumination variation, brightness adjustment, speckle noise injection, blur distortion, and structural anomalies. The assessment has been performed on the three subsets of the data to evaluate both robustness and generalization through unseen distortions.

On these samples, the following configurations are evaluated for each model: zero-shot on KITTI (COCO-trained), fine-tuned 10 epochs on clean KITTI, and fine-tuned 100 epochs on clean KITTI. The resulting per-sample $mAP$ values quantify not only domain transfer (COCO to KITTI) but also robustness to unseen corruptions that were not explicitly seen during training.

### 5.3 Image based Selection Model

Building upon the complementary behaviors observed between YOLO11 and RT-DETR, a selection model was developed which is the core contribution. Unlike the other methods of implementing a single model, which may have limited performance across conditions, the proposed framework performs image-level model arbitration. For each image, four descriptive features are extracted: which are average brightness, blur score (via Laplacian variance), entropy (information content), and edge density (Canny-based edge ratio). These four features characterize quality of the scene and structural complexity. The main objective of this model is to determine which model is more suitable for a given scene and choose it, which help to maximize the perception system accuracy, making use of the strengths of both architectures. This adaptive strategy ensures that the quality of the perception system is not constrained by the limitations of a single model and instead yields a more resilient and context-aware detection pipeline which is suitable for autonomous driving systems.

During training, A lightweight classifier (a random forest classifier model) is trained on the feature vectors with the winner labels as supervision. The objective is to learn decision boundaries in feature space corresponding to regimes where one detector systematically outperforms the other (modeling the correlation between abilities of global attention, and dense local convolutional features with different levels of blur, entropy, and edge density). This stage uses the $mAP$ output from the two detection models and use it as ground truth and the input is the image features. As shown in Figure 3.

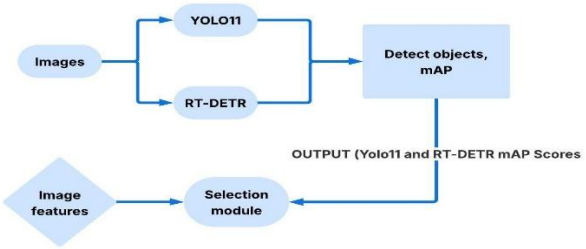


**Figure 3.** Architecture of the selection model during training phase

During Testing, once trained, the selection model is evaluated on new image folders (Samples 1–3) whose features are unseen during training. For each test image, the same four features are computed, the selection model predicts the preferred detector, and only that detector is applied to produce the final detections and $mAP$, as shown in Figure 4. Overall performance is then measured as $mAP_{(50-95)}$ of the selection-driven pipeline. This approach maintains high accuracy and performance while keeping low resource usage.

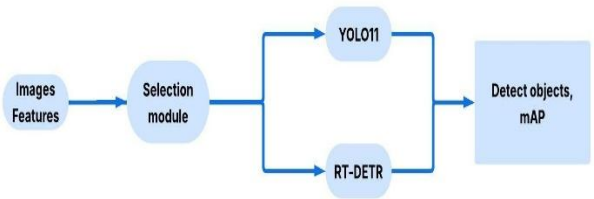


**Figure 4.** Architecture of the selection model during evaluation phase

## 6. RESULTS

The zero-shot results highlight a notable drop in performance due to domain shift. RT-DETR-L decreases from 53 $mAP_{(50-95)}$ on COCO to 34.3 on KITTI, while the version XL achieves 34.5. on the other hand, YOLO11-N achieves 44 $mAP_{(50-95)}$ on KITTI, and YOLO11-L reaches 51.3, highlighting the ability of YOLO11 versions to relatively generalize through domain shift without finetuning. These results indicate that convolutional inductive biases provide better structural alignment with automotive scenes than transformer attention patterns trained on generic imagery.

After fine-tuning on KITTI, both models show substantial improvement. RT-DETR-L enhanced to 61.7 $mAP_{(50-95)}$ after 10 epochs and achieves 79.1 after 100 epochs. On the other hand, YOLO11-L achieves 64.9 $mAP_{(50-95)}$ after 10 epochs and 76.2 after 100 epochs. These results indicates that, YOLO11

demonstrates faster early adaptation, on the other hand, RT-DETR surpasses YOLO after extended training, suggesting stronger long-term representational capacity.

YOLO11-L offers a higher starting point under zero-shot transfer, whereas RT-DETR-L surpasses YOLO11-L after extensive fine-tuning (79.1 vs. 76.2), indicating stronger long-term adaptation once sufficient target-domain data are seen. This complementarity demonstrated by higher initial robustness for YOLO and higher ceiling for RT-DETR motivates the integration via the selection model. Table 1 shows Summary of the results of the detection models.

| Models | Sample | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| RT-DETR-L (Zero-shot) | 33.4 | 32.5 | 31.1 |
| RT-DETR-XL (Zero-shot) | 33.5 | 32.4 | 31.4 |
| RT-DETR-L (Finetuned 10 epochs) | 52.9 | 46.7 | 46.9 |
| RT-DETR-L (Finetuned 100 epochs) | 61.1 | 51.8 | 56 |
| YOLO11-N (Zero-shot) | 39.8 | 36 | 36 |
| YOLO11-L (Zero-shot) | 49.1 | 44.9 | 46.2 |
| YOLO11-L (Finetuned 10 epochs) | 54 | 45.5 | 44.9 |
| YOLO11-L (Finetuned 100 epochs) | 60.4 | 48.6 | 51.6 |

**Table 1.** Summary of the results of the detection models

Across all three samples, both detectors suffer from the synthetic degradations relative to clean KITTI, but their relative ranking varies with training regime and corruption type: YOLO11 often retains stronger zero-shot performance, while RT-DETR gains more from fine-tuning and sometimes becomes more robust at high epochs under specific perturbations.

| RT-DETR | RT-DETR $\boldsymbol{mAP}$ | YOLO | YOLO $\boldsymbol{mAP}$ | Selection $\boldsymbol{mAP}$ |
|---|---|---|---|---|
| **Sample 1** | | | | |
| ZS-XL | 33.5 | FT-10 | 54.0 | **54.5** |
| FT-10 | 52.9 | ZS-L | 49.1 | **54.9** |
| FT-100 | 61.1 | ZS-L | 49.1 | **63.0** |
| **Sample 2** | | | | |
| ZS-L | 32.5 | FT-10 | 45.5 | **46.2** |
| FT-10 | 46.7 | ZS-L | 44.9 | **48.8** |
| FT-100 | 51.8 | ZS-L | 44.9 | **54.1** |
| **Sample 3** | | | | |
| ZS-L | 31.1 | FT-10 | 44.9 | **45.9** |
| FT-10 | 46.9 | ZS-L | 46.2 | **51.1** |
| FT-100 | 56.0 | ZS-L | 46.2 | **59.9** |

**Table 2.** Performance comparison of RT-DETR, YOLO, and the proposed selection model on sampled KITTI data, where ZS: Zero-shot, FT: Fine-tuned.

This non-uniform behaviour underlines the potential benefit of a conditional selection rather than committing to a single detector as the selection model achieves higher $mAP_{(50-95)}$ than the best individual model, with gains ranging from about 0.5 to nearly 4 points, depending on the sample and model pair. Importantly, these improvements are obtained without modifying the underlying detectors, purely by exploiting their complementarity through a learned, feature-based selection strategy, as shown in Table 2.

## 7. DISCUSSION

The experimental results provide several important insights into cross-domain generalization, model behavior, and the effectiveness of adaptive model selection in vision-based perception systems. First, the zero-shot evaluation highlights the significant impact of domain shift when transferring models from general-purpose datasets such as COCO to driving-specific datasets like KITTI. Both detectors experience notable performance degradation, with RT-DETR exhibiting a more pronounced drop compared to YOLO11. This can be attributed to differences in scene structure, object scale distribution, and viewpoint characteristics between the two datasets. The relatively stronger performance of YOLO11 under zero-shot conditions suggests that convolutional inductive biases provide better alignment with structured road scenes, enabling improved robustness without adaptation.

Second, fine-tuning experiments demonstrate that both architectures possess strong adaptation capabilities once exposed to target-domain data. While YOLO11 achieves faster performance gains during early-stage training, RT-DETR benefits more significantly from extended training and ultimately surpasses YOLO11 in peak accuracy. This behavior reflects the fundamental differences between the two architectures: convolutional models rely on localized feature extraction and converge quickly, whereas transformer-based models leverage global context modeling, which requires more data and training time to fully exploit.

Third, robustness analysis under synthetic degradations reveals that the relative performance of the two models is highly condition dependent. YOLO11 maintains more stable performance under moderate distortions, likely due to its reliance on local texture and edge features. In contrast, RT-DETR demonstrates improved robustness under more complex degradations after sufficient fine-tuning, suggesting that global attention mechanisms can better capture structural relationships when properly trained. These findings confirm that no single model consistently dominates across all conditions.

This observation motivates the proposed selection-based framework. Rather than relying on a fixed detector, the selection model enables dynamic adaptation to varying visual conditions by choosing the most suitable model for each input image. The performance gains observed across all degraded samples without modifying the underlying detection architectures, demonstrate that exploiting model complementarity is an effective and computationally efficient strategy.

An important advantage of the proposed approach is its efficiency. Unlike ensemble methods that require running multiple detectors in parallel, the selection model operates as a lightweight pre-inference module, ensuring that only a single detector is executed per image. This makes the framework particularly suitable for real-time autonomous driving systems,

where computational resources and latency are critical constraints.

From a broader perspective, the results highlight a shift from static perception pipelines toward adaptive perception systems. In real-world driving environments, visual conditions are highly dynamic and unpredictable. A context-aware system that can adapt its behavior based on scene characteristics is therefore more likely to maintain reliable performance. The proposed selection model represents a step in this direction by introducing a simple yet effective mechanism for adaptive decision-making in perception.

However, several limitations should be noted. The selection model relies on a small set of handcrafted image features, which may not fully capture all factors influencing detector performance. Additionally, the evaluation is conducted on a single dataset (KITTI), which, although representative, does not cover the full diversity of real-world driving conditions. Future work should explore more expressive feature representations, potentially incorporate learned features or temporal information, and validate the approach on larger and more diverse datasets.

## 8. CONCLUSION

This paper investigated the challenges of domain shift and robustness in vision-based object detection for autonomous driving. Through a systematic evaluation of YOLO11 and RT-DETR, we demonstrated that models pretrained on large-scale generic datasets experience significant performance degradation when transferred to driving-specific environments such as KITTI, with performance drops exceeding 15–20 $mAP$ points in zero-shot settings. However, both architectures exhibit strong adaptation capabilities through fine-tuning, reaching up to 79.1 $mAP$ for RT-DETR and 76.2 $mAP$ for YOLO11 after extended training, while revealing complementary behaviors in terms of generalization, convergence, and robustness.

Building on these observations, this work introduced an image-based selection model that chooses the most suitable detector for each input image based on simple yet effective visual descriptors. The proposed approach leverages the complementary strengths of convolutional and transformer-based detectors without modifying their architectures. Experimental results show that the selection strategy outperforms individual models, achieving improvements of up to 4 $mAP$ points across degraded scenarios, while maintaining computational efficiency by avoiding redundant inference.

The proposed framework highlights the importance of moving beyond static perception pipelines toward adaptive, context-aware systems capable of operating reliably under diverse and changing environments. By introducing a lightweight and interpretable mechanism for model selection, this work provides a practical pathway for improving perception robustness in real-world autonomous driving applications.

Future work will focus on extending the selection framework to incorporate additional detection architectures, exploring learned feature representations and temporal cues, and validating the approach on larger and more diverse datasets to further enhance generalization and scalability.

## ACKNOWLEDGEMENTS

This research has been supported by funding of Prof. Naser El-Sheimy from NSERC CREATE and Canada Research Chairs programs.

## REFERENCES

Arnold, E., Al-Jarrah, O.Y., Dianati, M., Fallah, S., Oxtoby, D., Mouzakitis, A., 2019. A Survey on 3D Object Detection Methods for Autonomous Driving Applications. IEEE Trans. Intell. Transport. Syst. 20, 3782–3795. https://doi.org/10.1109/TITS.2019.2892405

Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.-L., 2022. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA, pp. 1080–1089. https://doi.org/10.1109/CVPR52688.2022.00116

Chen, Y.-N., Dai, H., Ding, Y., 2022. Pseudo-Stereo for Monocular 3D Object Detection in Autonomous Driving, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA, pp. 877–887. https://doi.org/10.1109/CVPR52688.2022.00096

Chu, X., Deng, J., Li, Y., Yuan, Z., Zhang, Yanyong, Ji, J., Zhang, Yu, 2021. Neighbor-Vote: Improving Monocular 3D Object Detection through Neighbor Distance Voting, in: Proceedings of the 29th ACM International Conference on Multimedia. Presented at the MM ’21: ACM Multimedia Conference, ACM, Virtual Event China, pp. 5239–5247. https://doi.org/10.1145/3474085.3475641

Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 32(11), pp.1231–1237.

Gragnaniello, D., Greco, A., Saggese, A., Vento, M., Vicinanza, A., 2023. Benchmarking 2D Multi-Object Detection and Tracking Algorithms in Autonomous Vehicle Driving Scenarios. Sensors 23, 4024. https://doi.org/10.3390/s23084024

Guo, X., Shi, S., Wang, X., Li, H., 2021. LIGA-Stereo: Learning LiDAR Geometry Aware Representations for Stereo-based 3D Detector, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp. 3133–3143. https://doi.org/10.1109/ICCV48922.2021.00314

He, X., Yang, F., Yang, K., Lin, J., Fu, H., Wang, M., Yuan, J., Li, Z., 2024. SSD-MonoDETR: Supervised Scale-Aware Deformable Transformer for Monocular 3D Object Detection. IEEE Trans. Intell. Veh. 9, 555–567. https://doi.org/10.1109/TIV.2023.3311949

Jiang, Y., Zhang, L., Miao, Z., Zhu, X., Gao, J., Hu, W., Jiang, Y.-G., 2023a. PolarFormer: Multi-Camera 3D Object Detection with Polar Transformer. AAAI 37, 1042–1050. https://doi.org/10.1609/aaai.v37i1.25185

Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z., 2023. BEVDepth: Acquisition of Reliable Depth for Multi-View 3D Object Detection. AAAI 37, 1477–1485. https://doi.org/10.1609/aaai.v37i2.25233

Liu, Y., Wang, L., Liu, M., 2021. YOLOStereo3D: A Step Back to 2D for Efficient Stereo 3D Detection, in: 2021 IEEE International Conference on Robotics and Automation (ICRA). Presented at the 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, Xi’an, China, pp. 13018–13024. https://doi.org/10.1109/ICRA48506.2021.9561423

Park, D., Ambrus, R., Guizilini, V., Li, J., Gaidon, A., 2021. Is Pseudo-Lidar needed for Monocular 3D Object detection?, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp. 3122–3132. https://doi.org/10.1109/ICCV48922.2021.00313

Peng, X., Zhu, X., Wang, T., Ma, Y., 2022. SIDE: Center-based Stereo 3D Detector with Structure-aware Instance Depth Estimation, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Presented at the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, pp. 225–234. https://doi.org/10.1109/WACV51458.2022.00030

Roh, W., Chang, G., Moon, S., Nam, G., Kim, C., Kim, Y., Kim, J., Kim, S., 2023. ORA3D: Overlap Region Aware Multi-view 3D Object Detection. https://doi.org/10.48550/arXiv.2207.00865

Tardioli, D., Matellán, V., Heredia, G., Silva, M.F., Marques, L. (Eds.), 2023. ROBOT2022: Fifth Iberian Robotics Conference: Advances in Robotics, Volume 1, Lecture Notes in Networks and Systems. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-21065-5

Wang, S., Zheng, J., 2023. MonoSKD: General Distillation Framework for Monocular 3D Object Detection via Spearman Correlation Coefficient. https://doi.org/10.48550/arXiv.2310.11316

Xu, J., Peng, L., Chen, H., Li, H., Qian, W., Li, K., Wang, W., Cai, D., 2023. MonoNeRD: NeRF-like Representations for Monocular 3D Object Detection, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Paris, France, pp. 6791–6801. https://doi.org/10.1109/ICCV51070.2023.00627

Xu, J., Peng, L., Cheng, H., Xia, L., Zhou, Q., Deng, D., Qian, W., Wang, W., Cai, D., 2024. Regulating Intermediate 3D Features for Vision-Centric Autonomous Driving. AAAI 38, 6306–6314. https://doi.org/10.1609/aaai.v38i6.28449

Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao, Y., Lu, L., Zhou, J., Dai, J., 2023. BEVFormer v2: Adapting Modern Image Backbones to Bird’s-Eye-View Recognition via Perspective Supervision, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Vancouver, BC, Canada, pp. 17830–17839. https://doi.org/10.1109/CVPR52729.2023.01710

Yao, S., Guan, R., Huang, X., Li, Z., Sha, X., Yue, Yong, Lim, E.G., Seo, H., Man, K.L., Zhu, X., Yue, Yutao, 2024. Radar-Camera Fusion for Object Detection and Semantic Segmentation in Autonomous Driving: A Comprehensive Review. IEEE Trans. Intell. Veh. 9, 2094–2128. https://doi.org/10.1109/TIV.2023.3307157

Yoneda, K., Suganuma, N., Yanase, R., Aldibaja, M., 2019. Automated driving recognition technologies for adverse weather conditions. IATSS Research 43, 253–262. https://doi.org/10.1016/j.iatssr.2019.11.005

Zhang, Z.; Zhang, Z.; Li, G.; Xia, C., 2025. ZZ-YOLOv11: A lightweight vehicle detection model based on improved YOLOv11. Sensors (Basel), 25(11), 3399. doi:10.3390/s25113399.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J., 2023. DETRs beat YOLOs on real-time object detection. *arXiv preprint* arXiv:2304.08069.

Zhou, Q., Cao, J., Leng, H., Yin, Y., Kun, Y., Zimmermann, R., 2024. SOGDet: Semantic-Occupancy Guided Multi-View 3D Object Detection. AAAI 38, 7668–7676. https://doi.org/10.1609/aaai.v38i7.28600.