Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

# Enhancing Visual Perception A Deep Learning Approach to Object Detection and Segmentation in Real-World Environments

**Muhammad Faheem Khalid** (Corresponding Author)*
Masters In Data Science School Of Physics, Engineering And Computer Science, University Of Hertfordshire, Hatfield, United Kingdom
Email: faheemit10@gmail.com

**Waseem Akram**
Diploma in Computer Science, Intermediate from GCT, Lahore, Pakistan
Email: waseem261648@gmail.com

**Bisma Javed**
University: Department of Computer science, University of Narowal, Narowal, Pakistan. Email: bisma.javed440@gmail.com

## ABSTRACT

**Background:** Object detection and segmentation in unconstrained natural scenes remain among the most persistent challenges in computer vision. Despite significant advances in deep learning, performance under real-world conditions involving partial occlusion, cluttered backgrounds, and inconsistent illumination continues to demand further investigation.

**Aim:** This study systematically evaluates and benchmarks three state-of-the-art deep learning architectures, YOLOv8, Mask DINO, and Dilate Former with respect to detection precision, segmentation quality, computational efficiency, and robustness against visual degradation in complex real-world environments.

**Method:** A structured comparative experimental design was employed across three benchmark datasets: COCO, ADE20K, and Cityscapes, with 5,000 randomly sampled images from each. All models were trained under an identical pipeline incorporating data augmentation, model-specific hyperparameter tuning, and optimization via SGD or Adam algorithms. Evaluation covered five dimensions: mean Average Precision (mAP), mean Intersection over Union (mIoU), Panoptic Quality (PQ), per-frame inference latency, and occlusion robustness.

**Results:** Dilate Former outperformed competing models in segmentation, achieving an average mIoU of 0.83, PQ of 0.81, with less than 3% occlusion degradation and ±3% cross-dataset consistency. Mask DINO led in detection with the highest average mAP of 0.75 and mIoU of 0.77. YOLOv8 demonstrated the fastest inference at 18ms per frame, while recording an average mAP of 0.72 and mIoU of 0.66, with approximately 7% occlusion sensitivity.

**Conclusion:** Hybrid CNN-transformer architectures offer significant advantages for real-world visual perception tasks. Dilate Former suits precision-critical applications, while YOLOv8 is ideal for real-time deployments. Future work should explore

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

multimodal fusion and edge-optimized lightweight models.

**Keywords:** Object Detection, Image Segmentation, Yolov8, Mask Dino, Dilated Convolutions, Vision Transformers, Mean Average Precision, Intersection Over Union, Real-Time Inference, Deep Neural Networks

## INTRODUCTION

Visual perception in complex natural environments is a key challenge in computer vision. Accurate object detection, localization, and segmentation are crucial for applications like autonomous driving, medical imaging, surveillance, and robotics (LeCun, Bengio and Hinton, 2015). Traditional computer vision methods using handcrafted features like HOG and SIFT struggle with the variability and complexity of real-world scenes. They perform poorly under partial occlusion, illumination changes, and cluttered backgrounds (Dalal and Triggs, 2005).

The introduction of AlexNet by Krizhevsky, Sutskever and Hinton (2012) demonstrated that deep convolutional architectures trained on large-scale datasets could achieve unprecedented levels of image classification accuracy, fundamentally redefining the landscape of computer vision research. Subsequent developments, including the Region-based Convolutional Neural Network framework proposed by Girshick et al. (2014), established deep learning as the dominant paradigm for object detection tasks, enabling simultaneous object localization and classification with significantly improved accuracy over conventional methods.

Object detection identifies and localizes objects in an image using bounding boxes. Image segmentation extends this to pixel-level understanding, precisely delineating object boundaries and semantic regions (He et al., 2017). Instance segmentation distinguishes individual objects within the same category. This is crucial for fine-grained scene understanding in applications like autonomous driving and medical diagnostics (Bolya et al., 2019).

Object detection architectures have evolved through key milestones. Two-stage detectors, like Faster R-CNN, achieve high accuracy using region proposals followed by classification and bounding box refinement (Ren et al., 2015). Single-stage detectors, like the YOLO family, focus on computational efficiency by detecting objects in a single forward pass. This enables real-time inference for latency-sensitive applications (Redmon et al., 2016). The most recent iteration, YOLOv8, incorporates architectural refinements, including an anchor-free detection head and enhanced feature pyramid network, achieving competitive accuracy with exceptional inference speed (Jocher, Chaurasia and Qiu, 2023).

The advent of transformer-based architectures, originally proposed for natural language processing, has profoundly influenced the development of visual recognition models (Vaswani et al., 2017). The Vision Transformer uses self-attention to capture long-range spatial dependencies in images, complementing the local feature extraction of convolutional neural networks (Dosovitskiy et al., 2021). Mask DINO combines transformer-based detection and instance segmentation in a single framework, achieving state-of-the-art performance by jointly optimizing detection and segmentation objectives (Li et al., 2023). Dilate Former uses dilated attention within a transformer to capture multi-scale context, achieving superior performance in

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

semantic and panoptic segmentation (Chen et al., 2023).

Despite substantial progress, the deployment of deep learning models in real-world environments continues to present significant challenges. Performance degradation under partial occlusion, where target objects are partially obscured by other scene elements, remains a critical limitation of existing detection and segmentation frameworks (Wang et al., 2020). Similarly, domain shift across diverse datasets and environmental conditions poses considerable challenges to model generalization, limiting the practical applicability of models trained on controlled benchmark datasets to unconstrained real-world scenarios (Zhao et al., 2021). The trade-off between segmentation accuracy and computational efficiency further constrains the deployment of high-performance models in resource-limited environments such as edge computing platforms and embedded systems (Howard et al., 2019).

## METHODOLOGY

This study employed a structured comparative experimental framework to assess the performance of three prominent deep learning architectures, YOLOv8, Mask DINO, and Dilate Former for object detection and segmentation in complex real-world environments. These architectures were selected based on their state-of-the-art performance and complementary design philosophies, combining convolutional neural network efficiency with transformer-based contextual reasoning. The primary evaluation dimensions included detection precision, segmentation quality, computational efficiency, and robustness under challenging visual conditions.

### Datasets

Experiments were conducted across three widely adopted benchmark datasets: COCO, ADE20K, and Cityscapes. To ensure a balanced and representative evaluation, 5,000 images were randomly sampled from each dataset, collectively covering diverse real-world conditions including partial occlusion, densely cluttered scenes, and significant illumination variation. These datasets collectively provide comprehensive coverage of semantic, instance, and panoptic segmentation scenarios, enabling rigorous cross-domain generalization analysis.

### Training Pipeline

All three models were subjected to an identical and reproducible training pipeline to ensure experimental fairness. Input images were resized to a standardized resolution and normalized prior to training. A consistent set of data augmentation strategies was applied, including random horizontal flipping, random cropping, spatial scaling transformations, and color jittering, to enhance model generalization across diverse visual conditions. Each model underwent independent hyperparameter optimization, with training conducted using either Stochastic Gradient Descent or Adam optimization algorithms, supported by adaptive learning rate scheduling. Training was conducted for a maximum of 100 epochs with early stopping applied upon validation convergence. All experiments were conducted on an NVIDIA RTX 3090 GPU with 24GB VRAM to ensure consistency and reproducibility across all evaluations.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

**Evaluation Metrics**

Model performance was quantitatively assessed using five standardized evaluation metrics. Mean Average Precision was employed to measure object detection accuracy. Mean Intersection over Union and Panoptic Quality were utilized to evaluate segmentation performance at pixel and instance levels respectively. Per-frame inference latency in milliseconds was recorded to assess computational efficiency. Additionally, mIoU degradation under partial occlusion conditions was measured to evaluate model robustness in visually challenging scenarios. All experiments were executed on GPU-accelerated computational environments with identical hardware configurations to ensure consistency and reproducibility across all evaluations.

**Cross-Dataset Evaluation**

To assess the generalization capability of each architecture beyond its primary training domain, cross-dataset evaluation was performed. Each model trained on one dataset was subsequently evaluated on the remaining datasets, enabling a comprehensive analysis of domain adaptability and performance consistency across varied environmental conditions.

## RESULTS

The experimental evaluation was conducted using 5,000 randomly sampled images from each of the three benchmark datasets — COCO, ADE20K, and Cityscapes. The performance of YOLOv8, Mask DINO, and Dilate Former was systematically assessed across five evaluation dimensions: detection accuracy, segmentation quality, panoptic quality, inference speed, and robustness under occlusion.

**Detection Performance (mAP)**

Mask DINO achieved the highest detection accuracy with an average mAP of 0.75 across all datasets, demonstrating its superior capability in localizing and classifying objects in complex scenes. Dilate Former recorded a competitive average mAP of 0.74, while YOLOv8 achieved 0.72, reflecting a moderate trade-off between detection precision and inference efficiency.

Table 1: The object detection results measured using mean Average Precision across all three datasets.

| Model | COCO | ADE20K | Cityscapes | Avg mAP |
|---|---|---|---|---|
| YOLOv8 | 0.74 | 0.70 | 0.72 | 0.72 |
| Mask DINO | 0.77 | 0.74 | 0.75 | **0.75** |
| Dilate Former | 0.76 | 0.73 | 0.74 | 0.74 |

**Segmentation Performance (mIoU)**

Dilate Former demonstrated state-of-the-art segmentation performance, achieving the highest average mIoU of 0.83 across all benchmark datasets. This result confirms the effectiveness of dilated convolution mechanisms in capturing fine-grained spatial features necessary for precise pixel-level segmentation. Mask DINO recorded a strong average mIoU of 0.77, while YOLOv8 achieved 0.66, reflecting its primary

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

optimization toward detection speed rather than segmentation fidelity.

Table 2: The segmentation accuracy of each model measured using mean Intersection over Union.

| Model | COCO | ADE20K | Cityscapes | Avg mIoU |
|---|---|---|---|---|
| YOLOv8 | 0.68 | 0.65 | 0.66 | 0.66 |
| Mask DINO | 0.79 | 0.76 | 0.77 | 0.77 |
| Dilate Former | 0.84 | 0.82 | 0.83 | **0.83** |

**Panoptic Quality (PQ)**

Dilate Former outperformed competing architectures in Panoptic Quality, recording an average PQ of 0.81, confirming its superior unified detection and segmentation capabilities. Mask DINO achieved a competitive average PQ of 0.76, while YOLOv8 recorded 0.63, indicating reduced performance in complex panoptic scene understanding tasks.

Table 3: The Panoptic Quality scores, which collectively measure combined detection and segmentation performance

| Model | COCO | ADE20K | Cityscapes | Avg PQ |
|---|---|---|---|---|
| YOLOv8 | 0.64 | 0.61 | 0.63 | 0.63 |
| Mask DINO | 0.78 | 0.75 | 0.76 | 0.76 |
| Dilate Former | 0.82 | 0.80 | 0.81 | **0.81** |

**Inference Speed**

YOLOv8 demonstrated exceptional computational efficiency, achieving an inference speed of 18 milliseconds per frame, making it highly suitable for latency-sensitive real-time applications. Mask DINO and Dilate Former recorded higher inference latencies of 45ms and 55ms respectively, reflecting the computational overhead associated with their transformer-based architectural components, though this trade-off is justified by their substantially superior segmentation accuracy.

Table 4: The per-frame inference latency recorded for each model under identical hardware configurations.

| Model | Inference Time (ms/frame) |
|---|---|
| YOLOv8 | **18** |
| Mask DINO | 45 |
| Dilate Former | 55 |

**Occlusion Robustness**

Dilate Former demonstrated exceptional occlusion resilience, maintaining consistent segmentation performance with a degradation of less than 3% mIoU under partial occlusion scenarios. Mask DINO experienced a moderate performance reduction of approximately 4%, while YOLOv8 exhibited the highest sensitivity to occlusion with approximately 7% mIoU degradation, highlighting its limitations in visually challenging real-world conditions.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 5: The robustness of each model under partial occlusion conditions was evaluated by measuring mIoU degradation relative to non-occluded baseline performance.

| Model | mIoU Drop Under Occlusion | Robustness Rating |
|---|---|---|
| Dilate Former | < 3% | Excellent |
| Mask DINO | ~ 4% | Good |
| YOLOv8 | ~ 7% | Moderate |

**Cross-Dataset Generalization**

Dilate Former exhibited the strongest cross-dataset generalization capability, maintaining performance consistency within ±3% variation across all datasets. Mask DINO demonstrated a moderate generalization drop of approximately 4%, while YOLOv8 experienced a performance reduction of approximately 6%, suggesting relatively higher sensitivity to domain shift across diverse environmental conditions.

Table 6: Cross-dataset generalization was evaluated by measuring performance variation when models were tested beyond their primary training domain.

| Model | Performance Variation | Generalization Rating |
|---|---|---|
| Dilate Former | ± 3% | Excellent |
| Mask DINO | ~ 4% drop | Good |
| YOLOv8 | ~ 6% drop | Moderate |

## DISCUSSION

The findings of this study provide meaningful insights into the comparative capabilities of YOLOv8, Mask DINO, and Dilate Former for object detection and segmentation in complex real-world environments. The results align with and extend existing literature, while also revealing important practical implications for architecture selection in diverse deployment scenarios.

Dilate Former consistently achieved superior segmentation performance across all benchmark datasets, recording the highest average mIoU of 0.83 and Panoptic Quality of 0.81. Dilated attention mechanisms significantly enhance the capacity of transformer architectures to capture multi-scale contextual dependencies, enabling more precise boundary delineation and pixel-level classification in semantically complex scenes (Liu et al., 2022). Furthermore, the exceptional occlusion robustness of Dilate Former, with less than 3% mIoU degradation under partial occlusion, confirms that transformer-based architectures with expanded receptive fields maintain stronger feature representations under partial visual obstruction compared to purely convolutional counterparts (Zhu et al., 2022).

Mask DINO demonstrated the strongest object detection performance with an average mAP of 0.75, reinforcing that unified detection and segmentation frameworks leveraging deformable attention mechanisms achieve superior localization accuracy by enabling adaptive spatial sampling across feature maps (Zhang et al., 2022). The competitive segmentation performance of Mask DINO with an average mIoU of 0.77 further validates the architectural advantage of jointly optimizing detection and segmentation objectives within a single transformer-based framework (Hu et al.,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

2023).

The inference efficiency of YOLOv8 at 18 milliseconds per frame represents a significant practical advantage for real-time deployment scenarios, as anchor-free single-stage detectors achieve substantial reductions in computational overhead without proportional sacrifices in detection accuracy (Hussain, 2023). However, the comparatively lower segmentation accuracy of YOLOv8 with an average mIoU of 0.66 and higher occlusion sensitivity of approximately 7% mIoU degradation suggest inherent architectural limitations in capturing fine-grained spatial features necessary for precise segmentation in densely cluttered scenes (Kirillov et al., 2023).

The cross-dataset generalization results further highlight an important distinction among the evaluated architectures. Dilate Former maintained performance consistency within ±3% variation across datasets, while YOLOv8 exhibited approximately 6% performance drop, suggesting greater sensitivity to domain shift. Transformer-based architectures with global self-attention mechanisms develop more transferable feature representations compared to convolutional architectures, which tend to overfit domain-specific low-level visual patterns (Zhou et al., 2022). The moderate generalization drop observed in Mask DINO of approximately 4% reflects a balanced trade-off between architectural complexity and domain adaptability, consistent with generalization patterns observed in hybrid CNN-transformer detection frameworks evaluated across heterogeneous benchmark datasets (Xie et al., 2021).

## CONCLUSION

This study conducted a comprehensive comparative evaluation of YOLOv8, Mask DINO, and Dilate Former for object detection and segmentation across three benchmark datasets COCO, ADE20K, and Cityscapes. The findings demonstrated that Dilate Former achieved superior segmentation performance with an average mIoU of 0.83 and Panoptic Quality of 0.81, maintaining exceptional occlusion resilience of less than 3% degradation and cross-dataset consistency within ±3% variation, confirming the effectiveness of dilated convolution mechanisms for precise pixel-level scene understanding. Mask DINO proved most effective for object detection with the highest average mAP of 0.75 and competitive mIoU of 0.77, while YOLOv8 excelled in computational efficiency at 18ms per frame, making it ideal for real-time deployments despite its comparatively lower segmentation accuracy. Overall, these results affirm that hybrid CNN-transformer architectures offer substantial advantages for real-world visual perception tasks, with architecture selection guided by the specific trade-off between accuracy, segmentation quality, and inference speed. Future research should explore multimodal data fusion strategies and lightweight model optimization for edge computing deployments in resource-constrained real-world applications.

## REFERENCE

Bolya, D., Zhou, C., Xiao, F. and Lee, Y.J. (2019) 'YOLACT: Real-time instance segmentation', Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9157–9166.

Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L. and Zhang, L. (2023) 'Dilate Former: Visual recognition with dilated transformer', Proceedings of the

Online ISSN 3007-3197
Print ISSN 3007-3189



IEEE/CVF International Conference on Computer Vision, pp. 6299–6309.

Dalal, N. and Triggs, B. (2005) 'Histograms of oriented gradients for human detection', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. and Houlsby, N. (2021) 'An image is worth 16x16 words: Transformers for image recognition at scale', International Conference on Learning Representations. Available at: https://arxiv.org/abs/2010.11929

Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) 'Rich feature hierarchies for accurate object detection and semantic segmentation', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587.

He, K., Gkioxari, G., Dollar, P. and Girshick, R. (2017) 'Mask R-CNN', Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969.

Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M. and Adam, H. (2019) 'Searching for MobileNetV3', Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324.

Hu, J., Cao, L., Lu, H., Zhang, S., Wang, Y., Li, K. and Cao, J. (2023) 'Joint multi-task learning for semantic segmentation and object detection', IEEE Transactions on Image Processing, 32(1), pp. 1456–1470.

Hussain, M. (2023) 'YOLO-v1 to YOLO-v8: The rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection', Machines, 11(7), p. 677.

Jocher, G., Chaurasia, A. and Qiu, J. (2023) Ultralytics YOLOv8. Available at: https://github.com/ultralytics/ultralytics

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'ImageNet classification with deep convolutional neural networks', Advances in Neural Information Processing Systems, 25, pp. 1097–1105.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L. and Girshick, R. (2023) 'Segment anything', Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', Nature, 521(7553), pp. 436–444.

Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M. and Shum, H.Y. (2023) 'Mask DINO: Towards a unified transformer-based framework for object detection and segmentation', Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3041–3050.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z. and Guo, B. (2022) 'Swin transformer: Hierarchical vision transformer using shifted windows', Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

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

Ren, S., He, K., Girshick, R. and Sun, J. (2015) 'Faster R-CNN: Towards real-time object detection with region proposal networks', Advances in Neural Information Processing Systems, 28, pp. 91–99.

Online ISSN 3007-3197
Print ISSN 3007-3189



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. and Polosukhin, I. (2017) 'Attention is all you need', Advances in Neural Information Processing Systems, 30, pp. 5998–6008.

Wang, X., Zhang, R., Kong, T., Li, L. and Shen, C. (2020) 'SOLOv2: Dynamic and fast instance segmentation', Advances in Neural Information Processing Systems, 33, pp. 17721–17732.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P. (2021) 'SegFormer: Simple and efficient design for semantic segmentation with transformers', Advances in Neural Information Processing Systems, 34, pp. 12077–12090.

Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. (2021) 'Pyramid scene parsing network', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890.

Zhang, F., Chen, Y., Li, Z., Hong, Z., Liu, J., Ma, F. and Shao, L. (2022) 'ACFNet: Attentional class feature network for semantic segmentation', Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6798–6807.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A. and Torralba, A. (2022) 'Semantic understanding of scenes through the ADE20K dataset', International Journal of Computer Vision, 127(3), pp. 302–321.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X. and Dai, J. (2022) 'Deformable DETR: Deformable transformers for end-to-end object detection', International Conference on Learning Representations. Available at: https://arxiv.org/abs/2010.04159