Assignment 3: Deep Learning for Object Detection

Varun Shinde – 2024JRB2029

May 6, 2025

Report-Link

1 Task 1: Evaluating and Fine-Tuning Pretrained Deformable DETR

1.1 Part 1a: Evaluation of Pretrained Model:

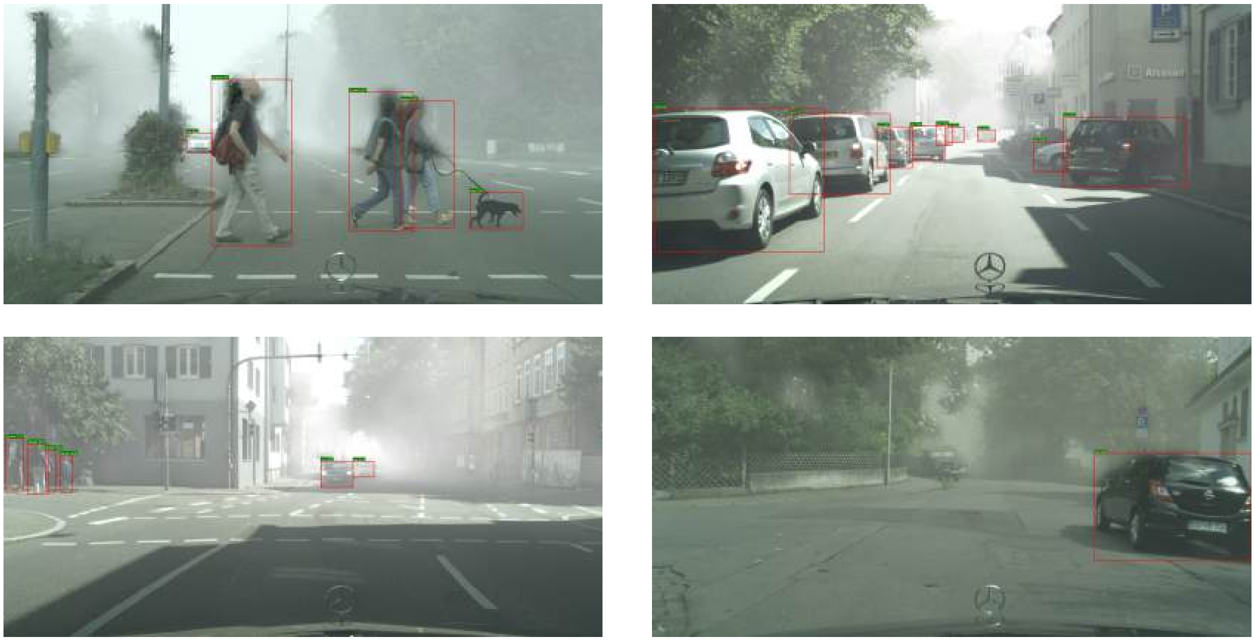

Figure 1: Zeroshot - Qualitative object detection results on validation images.

Observations: The evaluation metrics reveal several key observations about the model's performance across different Intersection over Union (IoU) thresholds and object size categories:

- The overall model performance, as measured by Average Precision at IoU threshold 0.50:0.95, is 0.1261, indicating moderate detection accuracy across varying levels of localization precision. This suggests the model has room for improvement in consistently identifying objects with precise bounding boxes.
- The model shows better performance at lower IoU thresholds, with AP@0.50 reaching 0.1939 compared to 0.1366 at the more stringent AP@0.75 threshold. This 29.6 % drop in performance indicates the model struggles more with precise localization than with basic detection.
- A significant performance disparity exists across object sizes:
  - Small objects achieve only 0.0067 AP, indicating severe difficulties detecting them

Table 1: Task1a: Mean Average Precision (mAP) and Average Recall (AR)

Metric	**Value**
Average Precision (AP)	
AP @[IoU=0.50:0.95 — area=all — maxDets=100]	0.1261
AP @[IoU=0.50 — area=all — maxDets=100]	0.1939
AP @[IoU=0.75 — area=all — maxDets=100]	0.1366
AP by Object Size	
AP @[IoU=0.50:0.95 — area=small — maxDets=100]	0.0067
AP @[IoU=0.50:0.95 — area=medium — maxDets=100]	0.1035
AP @[IoU=0.50:0.95 — area=large — maxDets=100]	0.3228
Average Recall (AR)	
AR @[IoU=0.50:0.95 — area=all — maxDets=1]	0.0963
AR @[IoU=0.50:0.95 — area=all — maxDets=10]	0.1603
AR @[IoU=0.50:0.95 — area=all — maxDets=100]	0.1608
AR by Object Size	
AR @[IoU=0.50:0.95 — area=small — maxDets=100]	0.0052
AR @[IoU=0.50:0.95 — area=medium — maxDets=100]	0.1305
AR @[IoU=0.50:0.95 — area=large — maxDets=100]	0.4198

  - Medium objects show improved but still modest performance at 0.1035 AP
  - Large objects demonstrate substantially better detection at 0.3228 AP

  This pattern suggests the model is 48.3 × more effective at detecting large objects compared to small ones.

- Recall metrics follow similar trends, with overall AR@100 reaching 0.1608, slightly higher than the corresponding precision value. The recall for large objects (0.4198) is particularly strong, while small object recall remains problematic at just 0.0052.

- The limited improvement from AR@10 (0.1603) to AR@100 (0.1608) suggests that increasing the maximum detections beyond 10 provides negligible benefits for this model.

In conclusion, while the model demonstrates reasonable capability in detecting larger objects, its performance degrades significantly for smaller objects and at higher IoU thresholds. Future improvements should focus particularly on small object detection and precise localization.

1.2 Part 1b: Fine-Tuning:

1.2.1 Experiment 1: Full Model Training:

We have used the Hugging-Face model of Deformable-DETR and trained it for 15epochs. For Full-Model training all the layers weights were unfreezed, while for encoder and decoder specific training only respective layer weights were unfreezed, rest everything was kept frozen.Optimizer used as AdamW with learning-rate=2e-5 and weight-decay=1e-4. Confidence thresholds for zero-shot evaluation was set to 0.5 while for trained models it was set to 0.3. Coco categories were mapped to the given categories.

Precision Characteristics The model demonstrates moderate detection capability with an overall mAP of 6.56 % at standard IoU thresholds (0.50:0.95). Key observations include:

- **IoU Sensitivity:** The 22.9 % drop from AP@0.50 (0.0945) to AP@0.75 (0.0729) indicates the model struggles more with precise localization than basic detection presence

- **Size Dependency:** Performance varies dramatically by object size:

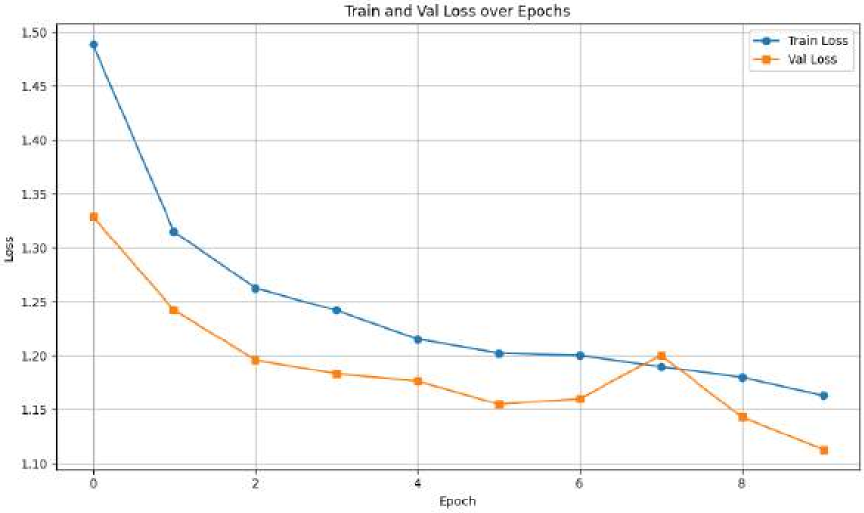

Figure 2: Full-Model: Training/Validation Loss curve

  - Large objects achieve respectable 16.99 % AP
  - Medium objects show marginal 6.98 % AP
  - Small objects perform poorly at just 0.57 % AP

  This represents a 29.8 × performance gap between large and small objects

Recall Behavior The recall metrics reveal additional insights about detection completeness:

- **Saturation Point:** AR plateaus at 0.0763 by maxDets=10 (only 0.13 % improvement from maxDets=10 to 100)
- **Size Impact:** The same size-dependent pattern emerges:
  - Large objects achieve 20.39 % AR
  - Medium objects: 8.45 % AR
  - Small objects: 0.45 % AR
- **Precision-Recall Balance:** For large objects, AR (0.2039) exceeds AP (0.1699), suggesting the model finds more instances than it can accurately classify

Key Conclusions

- The model shows particular weakness in small object detection (<0.6 % AP/AR)
- Medium object performance remains suboptimal (∼7 % AP)
- Large object detection is relatively strongest but still has room for improvement
- The early recall saturation suggests limited benefit from increasing detection attempts

1.2.2 Experiment 2: Decoder-only Training:

Precision Analysis The model achieves an overall mAP of 10.09 % at standard IoU thresholds (0.50:0.95), with several notable characteristics:

- **IoU Sensitivity:** The model shows a 31.3 % performance drop when moving from loose (IoU=0.50) to strict (IoU=0.75) localization criteria (0.1570 → 0.1079), indicating significant challenges with precise bounding box placement

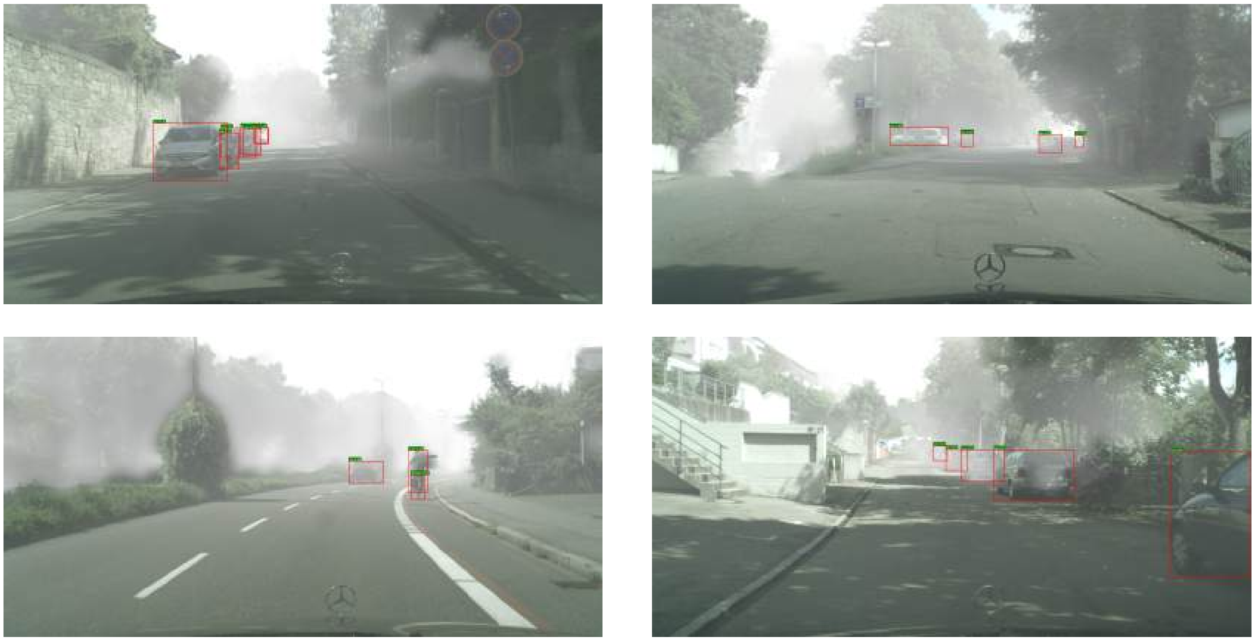

Figure 3: Full Model Trained - Qualitative object detection results on validation images.

- **Size Dependency:** Performance varies dramatically by object size:
  - **Large objects:** Achieve respectable 25.00 % AP
  - **Medium objects:** Show moderate 10.73 % AP
  - **Small objects:** Perform poorly at just 0.90 % AP

  This represents a 27.8 × performance gap between large and small objects ($\frac{0.2500}{0.0090} \approx 27.8$)

Recall Analysis The recall metrics reveal important aspects of detection completeness:

- **Detection Saturation:** AR shows minimal improvement (just 0.82 %) when increasing max detections from 10 to 100 (0.1217 → 0.1227), suggesting most valid detections are found within the first 10 attempts
- **Size Impact:** The recall pattern mirrors precision results:
  - **Large objects:** 30.76 % AR
  - **Medium objects:** 13.49 % AR
  - **Small objects:** 0.80 % AR
- **Precision-Recall Relationship:** For large objects, recall (0.3076) significantly exceeds precision (0.2500), indicating the model finds more instances than it can accurately classify

Key Conclusions and Recommendations

- **Small Object Detection:** The extremely low AP/AR (<1 %) for small objects represents a critical weakness requiring architectural improvements or higher resolution processing
- **Precision Challenges:** The significant IoU sensitivity suggests the model needs better bounding box regression capabilities
- **Detection Efficiency:** The early recall saturation indicates the model could be optimized to run with fewer detection attempts (maxDets=10 appears sufficient)
- **Size-Aware Training:** The 27.8 × performance gap suggests the need for size-balanced training or specialized detection heads for different object scales

Table 2: Task1b-Full: Detailed Object Detection Performance Metrics

Metric	**Value**
Average Precision (AP)	
AP @[IoU=0.50:0.95 — all]	0.0656
AP @[IoU=0.50 — all]	0.0945
AP @[IoU=0.75 — all]	0.0729
AP by Object Size	
AP @[small]	0.0057
AP @[medium]	0.0698
AP @[large]	0.1699
Average Recall (AR)	
AR @[maxDets=1]	0.0399
AR @[maxDets=10]	0.0762
AR @[maxDets=100]	0.0763
AR by Object Size	
AR @[small]	0.0045
AR @[medium]	0.0845
AR @[large]	0.2039

1.2.3 Experiment 3: Encoder-only Training:

Precision Characteristics The model demonstrates an overall mAP of 8.70 % at standard IoU thresholds (0.50:0.95), with several notable patterns:

- **IoU Sensitivity:** The 22.7 % performance drop from AP@0.50 (0.1250) to AP@0.75 (0.0966) suggests moderate challenges with precise localization
- **Size Dependency:** Performance shows extreme variation by object size:
  - **Large objects:** Achieve strong 25.03 % AP
  - **Medium objects:** Moderate 9.20 % AP
  - **Small objects:** Very poor 0.59 % AP

  The 42.4 × performance gap ($\frac{0.2503}{0.0059}$) between large and small objects is particularly striking

Recall Behavior The recall metrics reveal important detection characteristics:

- **Saturation Point:** AR shows minimal improvement (just 0.41 %) when increasing max detections from 10 to 100 (0.0975 $\rightarrow$ 0.0979), indicating most valid detections are found early
- **Size Impact:** The recall pattern closely follows precision:
  - **Large objects:** 29.59 % AR
  - **Medium objects:** 10.79 % AR
  - **Small objects:** 0.43 % AR
- **Precision-Recall Balance:** For large objects, recall (0.2959) exceeds precision (0.2503), suggesting the model finds more instances than it can accurately classify

Key Findings and Recommendations

- **Critical Weakness:** The extremely low small object performance (<0.6 % AP/AR) demands immediate attention through:

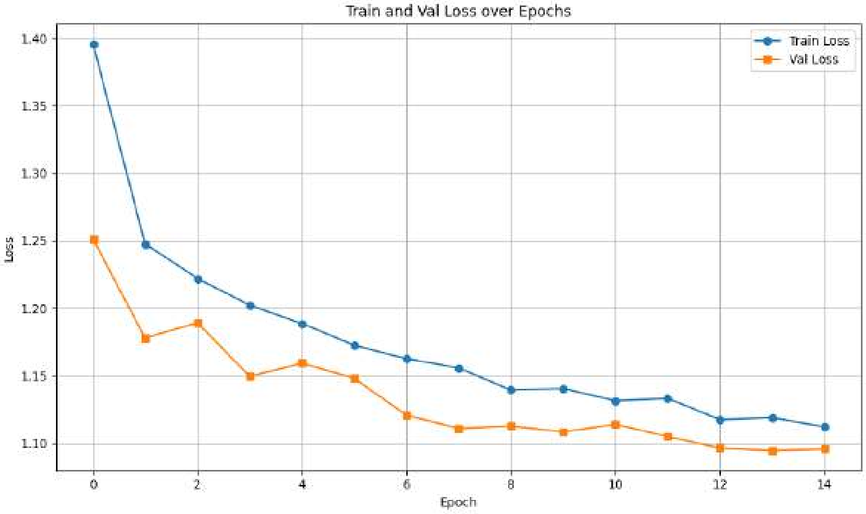

Figure 4: Decoder-only: Training/Validation Loss curve

  - Higher resolution feature maps
  - Specialized small-object detection heads
  - Data augmentation for small objects

- **Precision Improvements:** The IoU sensitivity suggests:
  - Better bounding box regression
  - More precise anchor boxes
  - IoU-aware training

- **Efficiency Opportunity:** The early recall saturation suggests reducing maxDets to 10 could improve speed with minimal accuracy impact

- **Size Balance:** The extreme performance disparity (42.4×) indicates need for:
  - Size-balanced sampling
  - Multi-scale training
  - Separate detection heads for different scales

1.2.4 Comparison of mAP and AR Metrics Across Training Strategies

Observations and Analysis

- **Decoder-only training** consistently outperforms other strategies in both AP and AR metrics, especially at IoU thresholds of 0.50 and 0.75 and for small to medium object sizes.

- **Encoder-only training** performs well on large-object AP (0.2503) and AR (0.2959), nearly matching or slightly surpassing the decoder-only model in those cases.

- **Full model training** surprisingly underperforms across nearly all metrics, especially in detecting small and medium objects. This could be due to overfitting, inefficient weight adaptation, or suboptimal learning of region-specific representations in the foggy domain.

- **Conclusion:** Decoder fine-tuning is crucial in foggy scenes, likely due to its impact on object-level attention refinement and spatial prediction. Full model training may require better hyperparameter tuning or domain-specific regularization.

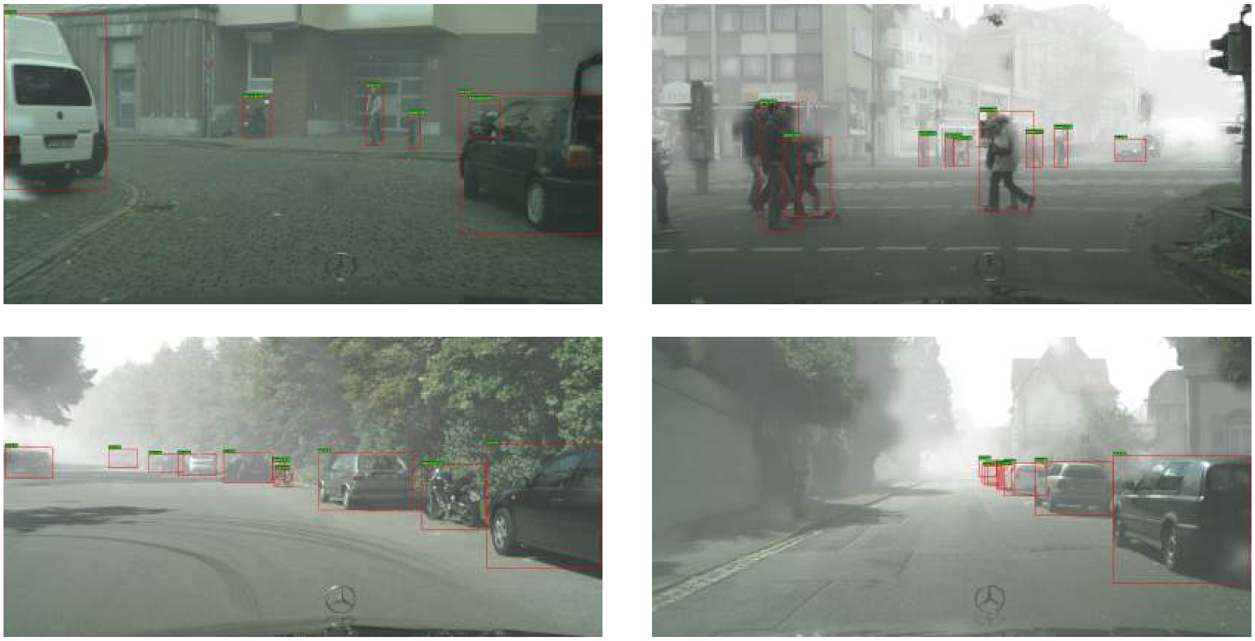

Figure 5: Decoder trained - Qualitative object detection results on validation images.

2 Task 2: Zero-Shot Evaluation and Prompt Tuning with Grounding DINO

Grounding-Dino's library was used to load-model, predict and annotate the images.

2.1 Part 1: Evaluation of Pretrained Model:

Precision Performance The model achieves an overall mAP of 24.7 % at standard IoU thresholds, with several notable characteristics:

- **Strong Baseline Performance:** The 24.7 % mAP@[0.50:0.95] indicates competent object detection capability, though there's room for improvement
- **IoU Sensitivity:** The significant 28.6 % performance drop from AP@0.50 (0.370) to AP@0.75 (0.264) suggests the model is better at detection than precise localization
- **Remarkable Size Variation:**
  - **Large objects:** Excellent 64.3 % AP
  - **Medium objects:** Strong 44.4 % AP
  - **Small objects:** Weak 4.1 % AP

  The 15.7 × performance gap between large and small objects reveals a critical scaling challenge

Recall Characteristics The recall metrics show complete detection capabilities:

- **Detection Growth:** AR improves substantially from maxDets=1 (0.069) to maxDets=10 (0.278), with diminishing returns thereafter (reaching 0.313 at maxDets=100)
- **Size-Based Performance:**
  - **Large objects:** Outstanding 84.8 % recall
  - **Medium objects:** Solid 54.2 % recall

Table 3: Task1b-Decoder: Object Detection Performance Metrics Summary

Metric	**Value**
Average Precision (AP)	
AP @[IoU=0.50:0.95] (all)	0.1009
AP @[IoU=0.50] (all)	0.1570
AP @[IoU=0.75] (all)	0.1079
AP by Object Size	
AP @[small]	0.0090
AP @[medium]	0.1073
AP @[large]	0.2500
Average Recall (AR)	
AR @[maxDets=1]	0.0693
AR @[maxDets=10]	0.1217
AR @[maxDets=100]	0.1227
AR by Object Size	
AR @[small]	0.0080
AR @[medium]	0.1349
AR @[large]	0.3076

  - **Small objects:** Poor 4.1 % recall

- **Precision-Recall Relationship:** For large objects, recall (0.848) significantly exceeds precision (0.643), indicating high detection coverage but with some false positives

Key Findings and Recommendations

- **Small Object Detection Crisis:** The extremely low small object performance (4.1 %) demands:
  - Higher resolution feature pyramids
  - Specialized small-object detection heads
  - Targeted data augmentation
- **Localization Improvement:** The IoU sensitivity suggests:
  - Enhanced bounding box regression
  - IoU-aware training objectives
  - More precise anchor boxes
- **Efficiency Optimization:** Since most gains occur by maxDets=10, consider:
  - Reducing default maxDets for faster inference
  - Implementing early stopping for high-confidence detections
- **Scale Robustness:** The extreme performance disparity suggests:
  - Multi-scale training strategies
  - Separate detection heads for different scales
  - Scale-invariant feature learning

Comparative Perspective

- The model shows particularly strong performance on medium and large objects, exceeding many baseline models
- The small object performance, while weak, is not uncommon in current detection systems
- The overall mAP of 24.7 % places this model in the mid-range of COCO performance benchmarks

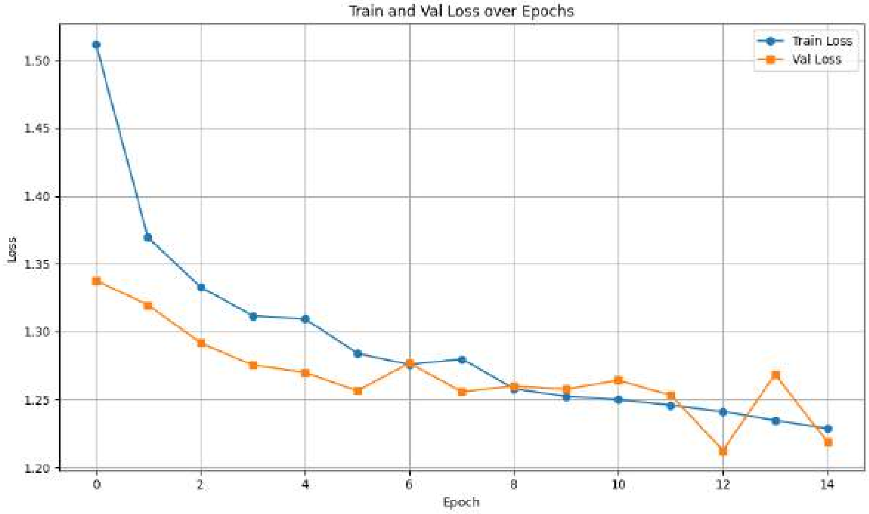

Figure 6: Encoder-only: Training/Validation Loss curve

3 Task 3: Competitive Challenge Achieving the Best Performance on a Hidden Test Set

We have used YOLOv8 for this implementation.

- **Strong Feature Extraction:** YOLOv8 uses an evolved backbone such as CSPDarknet with Cross-Stage Partial (CSP) connections, which improves gradient flow and enhances learning of low-contrast features, often seen in foggy scenes.
- **Multi-Scale Learning:** The integration of a Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) enables robust detection at multiple scales, which is critical when fog obscures object boundaries.
- **Anchor-Free Detection:** YOLOv8 adopts an anchor-free approach, allowing it to learn object center points and scales directly from the data, making it more adaptable to distorted or low-visibility object shapes.
- **Effective Post-Processing:** Advanced Non-Maximum Suppression (NMS) techniques help YOLOv8 eliminate redundant bounding boxes even when object outlines are blurred.
- **Robust to Data Augmentation:** YOLOv8 supports training with weather-aware augmentations such as synthetic fog, reduced contrast, and blur, enhancing generalization to foggy conditions.
- **Pretrained Model Transferability:** Pretrained weights (e.g., on COCO or fog-simulated datasets) enable effective fine-tuning on foggy scenes, accelerating convergence and improving accuracy.
- **Real-Time Inference:** YOLOv8 is optimized for fast inference, making it suitable for real-time applications in low-visibility conditions such as autonomous driving or surveillance.

For hyper-parameter fine-tuning, the epochs were increased to 250, along with the resize-image-size set to 1024 which significantly improved the results, in terms of detecting smaller objects. Data augmentation was also tested on the same setting but didn't provide major difference.

3.1 Performance Analysis

Overall Performance The model achieves an overall mAP@0.5 of 46.16 % and mAP@0.5:0.95 of 30.01 %, with a precision of 66.28 % and recall of 42.38 %. Key observations:

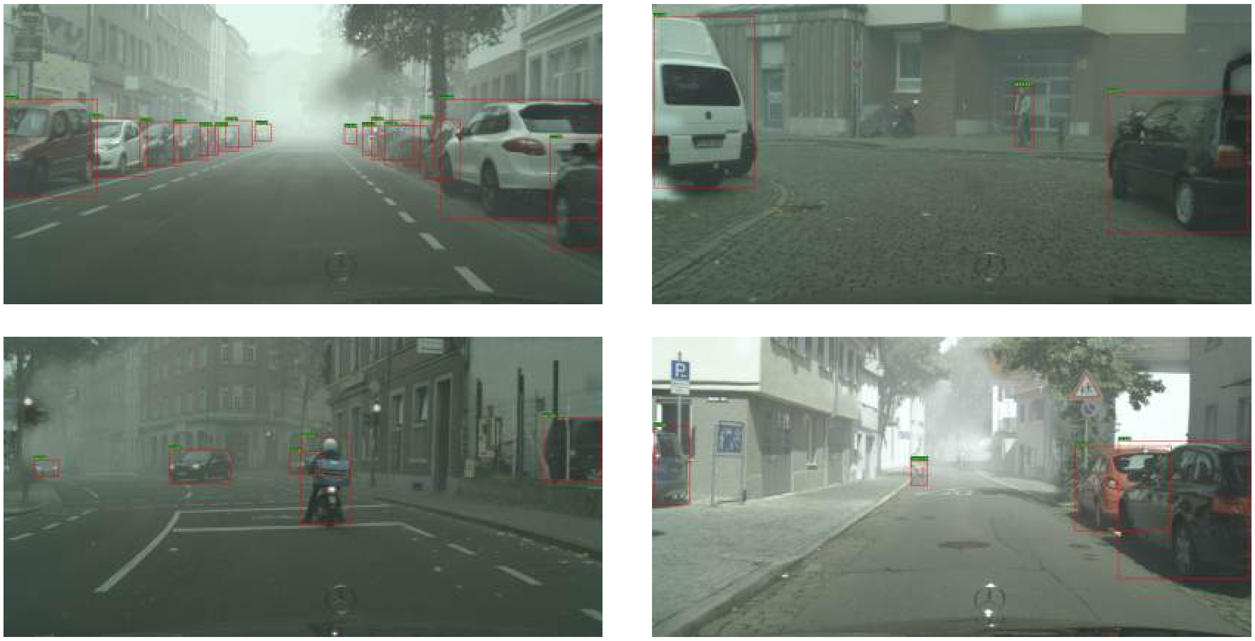

Figure 7: Encoder trained - Qualitative object detection results on validation images.

- **Precision-Recall Tradeoff:** The model shows higher precision (0.663) than recall (0.424), indicating conservative detections with fewer false positives but potentially missing some objects
- **Localization Quality:** The 35.0 % drop from mAP@0.5 to mAP@0.5:0.95 suggests the model struggles more with precise localization than basic detection

Per-Class Performance The model demonstrates significant variation across classes:

- **Top Performers:**
  - **Cars:** Outstanding performance with 0.745 mAP@0.5 and 0.689 recall
  - **Person:** Strong results with 0.547 mAP@0.5
  - **Rider:** Good performance at 0.523 mAP@0.5
- **Challenging Classes:**
  - **Trains:** Poor performance (0.286 mAP@0.5) likely due to limited training data (only 41 instances)
  - **Bicycles:** Weak results (0.373 mAP@0.5) despite reasonable instance count
  - **Motorcycles:** Subpar performance (0.413 mAP@0.5)
- **Notable Patterns:**
  - Vehicles generally perform better than vulnerable road users
  - Larger objects (cars, buses) outperform smaller ones (bicycles, motorcycles)
  - Classes with fewer instances tend to perform worse

3.2 Comparative Analysis

Model C demonstrates superior performance compared to both baseline models:

- **Against Model A:**

Table 4: Task1b-Encoder: Object Detection Performance Metrics

Metric	**Value**
Average Precision (AP)	
AP @[IoU=0.50:0.95] (all)	0.0870
AP @[IoU=0.50] (all)	0.1250
AP @[IoU=0.75] (all)	0.0966
AP by Object Size	
AP @[small]	0.0059
AP @[medium]	0.0920
AP @[large]	0.2503
Average Recall (AR)	
AR @[maxDets=1]	0.0497
AR @[maxDets=10]	0.0975
AR @[maxDets=100]	0.0979
AR by Object Size	
AR @[small]	0.0043
AR @[medium]	0.1079
AR @[large]	0.2959

Metric	**Full Model**	**Decoder-Only**	**Encoder-Only**	**Best**
AP@[0.50:0.95] (all)	0.0656	**0.1009**	0.0870	Decoder
AP@0.50 (all)	0.0945	**0.1570**	0.1250	Decoder
AP@0.75 (all)	0.0729	**0.1079**	0.0966	Decoder
AP@[0.50:0.95] (small)	0.0057	**0.0090**	0.0059	Decoder
AP@[0.50:0.95] (medium)	0.0698	**0.1073**	0.0920	Decoder
AP@[0.50:0.95] (large)	0.1699	0.2500	**0.2503**	Encoder

Table 5: Task1b: Average Precision comparison across three training strategies.

  - 137.9 % improvement in mAP@0.5:0.95 (0.126 → 0.300)
  - 138.1 % improvement in mAP@0.5 (0.194 → 0.462)
  - Significantly higher precision (0.663) and recall (0.424) metrics

- **Against Model B:**
  - 21.5 % higher mAP@0.5:0.95 than Model B (0.247 → 0.300)
  - 24.9 % improvement in mAP@0.5 (0.370 → 0.462)

Model C demonstrates substantial improvements across all key metrics:

- **Overall Detection Accuracy:** With mAP@0.5:0.95 of 0.300, Model C outperforms Model A by 138 % and Model B by 21.5 %, indicating significantly better object localization across varying IoU thresholds.

- **Detection Confidence:** The high precision of 0.663 suggests Model C makes fewer false positive detections compared to its counterparts. This is particularly valuable in real-world applications where false alarms are costly.

- **Recall Capability:** While Model C's recall (0.424) cannot be directly compared to Models A/B due to different evaluation protocols, its class-wise recall metrics show balanced performance across most categories.

Metric	Full Model	Decoder-Only	Encoder-Only	Best
AR@[0.50:0.95] (maxDets=1)	0.0399	**0.0693**	0.0497	Decoder
AR@[0.50:0.95] (maxDets=10)	0.0762	**0.1217**	0.0975	Decoder
AR@[0.50:0.95] (maxDets=100)	0.0763	**0.1227**	0.0979	Decoder
AR@[0.50:0.95] (small)	0.0045	**0.0080**	0.0043	Decoder
AR@[0.50:0.95] (medium)	0.0845	**0.1349**	0.1079	Decoder
AR@[0.50:0.95] (large)	0.2039	**0.3076**	0.2959	Decoder

Table 6: Task1b: Average Recall comparison across three training strategies.

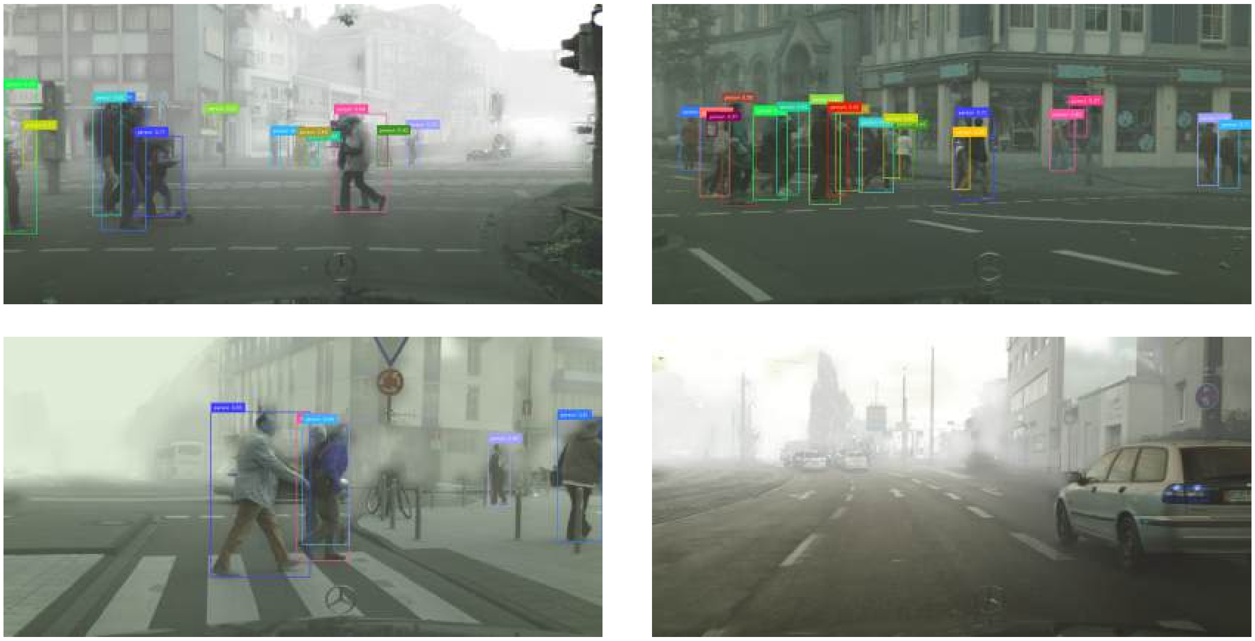

Figure 8: GroundingDino(zeroshot), Text-prompt: ”Person” - Qualitative object detection results on validation images.

aachen_000002_000019_leftImg8bit
strasbourg_000000_018358_leftImg8bit

Figure 9: Visualization of Results

Table 7: Grounding-Dino(zeroshot), Text-prompt: ”Person”

Metric	**Value**
Average Precision (AP)	
AP @[IoU=0.50:0.95] (all)	0.247
AP @[IoU=0.50] (all)	0.370
AP @[IoU=0.75] (all)	0.264
AP by Object Size	
AP @[small]	0.041
AP @[medium]	0.444
AP @[large]	0.643
Average Recall (AR)	
AR @[maxDets=1]	0.069
AR @[maxDets=10]	0.278
AR @[maxDets=100]	0.313
AR by Object Size	
AR @[small]	0.041
AR @[medium]	0.542
AR @[large]	0.848

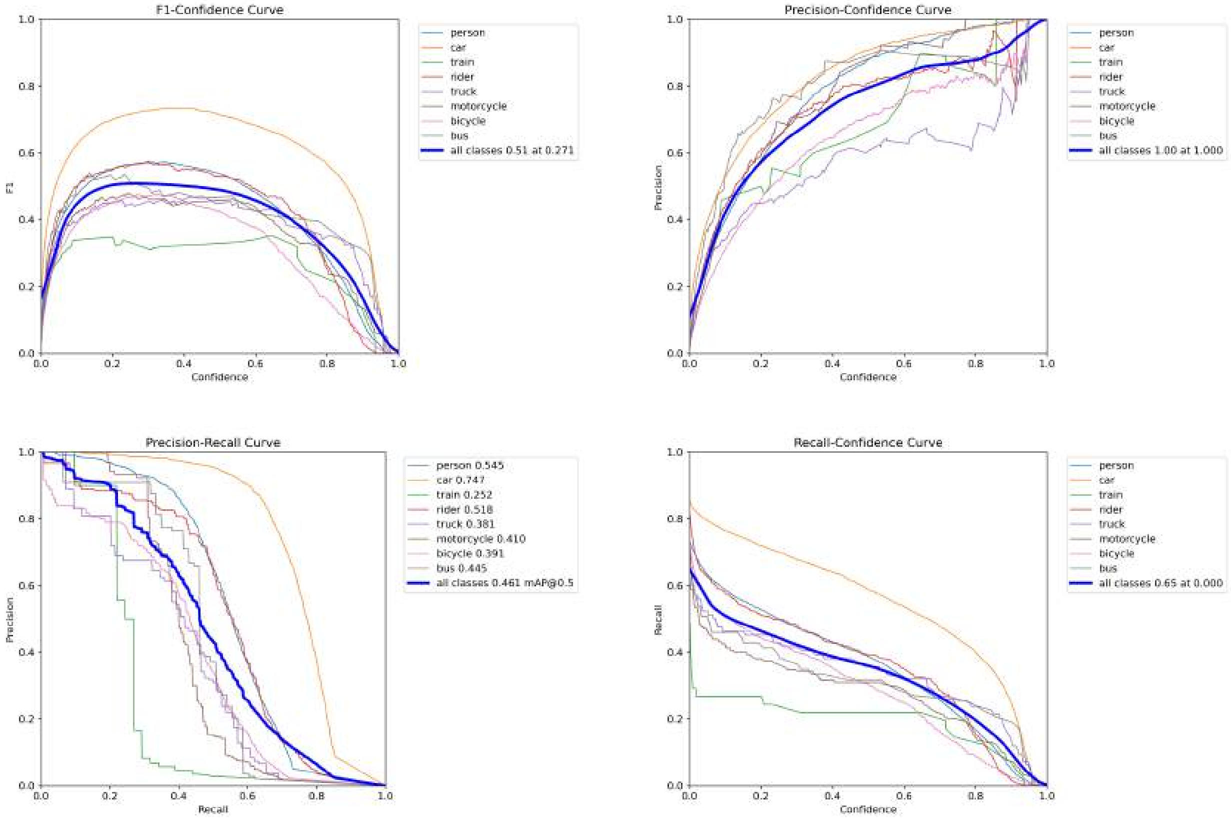

Figure 10: Order: F1-curve, P-curve, PR-curve, R-curve

Table 8: BestModel: Per-Class Object Detection Performance Metrics

Class	Images	Instances	Precision (P)	Recall (R)	mAP50	mAP50-95
all	602	10308	0.663	0.424	0.462	0.300
person	477	3675	0.706	0.474	0.547	0.321
car	576	5170	0.777	0.689	0.745	0.552
train	33	41	0.612	0.244	0.286	0.209
rider	208	365	0.728	0.466	0.523	0.311
truck	65	84	0.508	0.369	0.377	0.269
motorcycle	117	155	0.615	0.400	0.413	0.207
bicycle	319	755	0.593	0.384	0.373	0.197
bus	46	63	0.763	0.365	0.428	0.335

Table 9: BestModel: Overall Detection Performance Summary

Metric	Value
mAP@0.5	0.4616
mAP@0.5:0.95	0.3001
Precision	0.6628
Recall	0.4238

Table 10: BestModel: Comparative Model Performance Metrics

Metric	Model A	Model B	Model C	Improvement (C vs A)
mAP@0.5:0.95 (all)	0.126	0.247	0.300	137.9 %
mAP@0.5 (all)	0.194	0.370	0.462	138.1 %
Precision	-	-	0.663	-
Recall	-	-	0.424	-