COL780 Assignment - 3

Vidushi Maheshwari (2021CS10083)

May 8, 2025

Contents

Introduction

Introduction

Object detection is a fundamental task in computer vision that involves localizing and classifying multiple objects within an image. Recent advancements in deep learning have significantly improved the accuracy and efficiency of object detectors, especially with the introduction of transformer-based architectures such as DETR and its variants, and vision-language models like Grounding DINO. This assignment explores state-of-the-art object detection techniques through three major tasks.

- In Task 1, we evaluate and fine-tune a pre-trained Deformable DETR model on a subset of the Foggy Cityscapes dataset with COCO-style annotations. The goal is to analyze the effects of different fine-tuning strategies on detection performance, such as full model training versus selective training of encoder or decoder components.
- In Task 2, we examine the zero-shot detection capabilities of the Grounding DINO model, which combines visual features with language prompts for open-set object detection. We further explore prompt tuning strategies by introducing learnable prompt embeddings, aiming to improve detection quality in a few-shot or zero-shot setting.
- In Task 3, we participate in a competitive detection challenge, where we are free to choose any model architecture and optimization techniques to achieve the best possible performance on a hidden test set, using the validation set for evaluation and comparison.

Part I
Evaluating and Fine-Tuning Pretrained Deformable DETRs

§1 Subtask 1: Evaluation of Pretrained Model

- The code loads the pretrained `SenseTime/deformable-detr` model and image processor from the Hugging Face Transformers library. Also, loads the COCO-style ground truth annotations and constructs dictionaries mapping image IDs to file names and ground truth bounding boxes.
- For each image in the validation set, the image is read and pre-processed using the Hugging Face image processor. The model performs inference and outputs bounding boxes, class labels, and confidence scores. Predictions are converted to COCO format and stored for evaluation.
- All predictions are saved in a JSON file in COCO detection format. The `pycocotools` API is used to compute standard detection metrics such as mean Average Precision (mAP) across various IoU thresholds.

§1.1 Qualitative Analysis

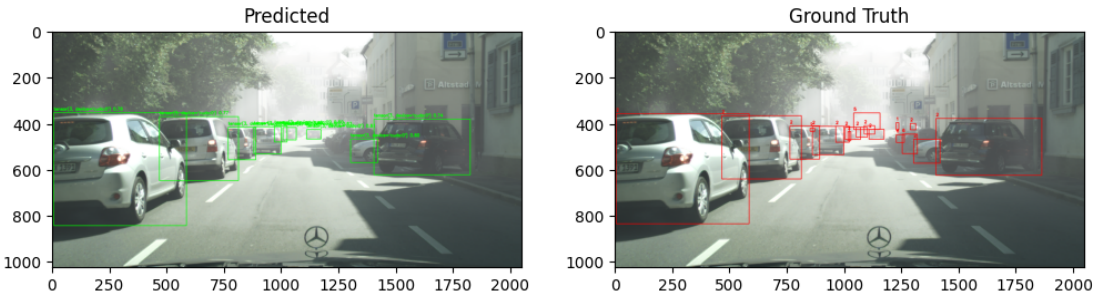


Image ID: 2050
Prediction: Detects many cars, including those further down the street, but again produces extra boxes not present in the ground truth.
Analysis: Good coverage of prominent vehicles with over-prediction and some missed detections in dense, foggy traffic scenes.

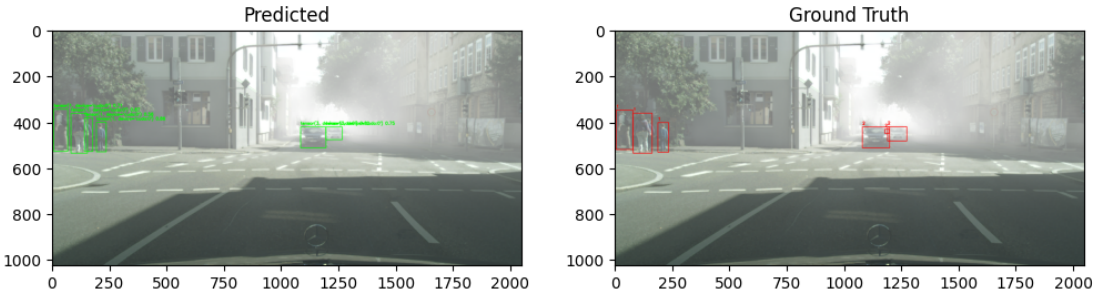


Image ID: 2060
Prediction: Multiple boxes are predicted, especially on the left, but most do not correspond to ground truth objects. Only 2 cars in the center is correctly detected.
Analysis: The model fails to detect pedestrians and produces many false positives, especially in cluttered or low-contrast areas.

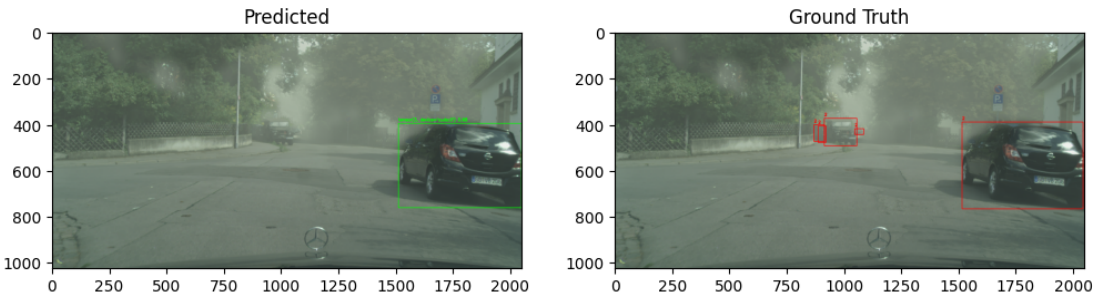


Image ID: 2080
Prediction: The car in the foreground is detected well, but the model misses smaller cars in the distance and does not predict boxes for them.
Analysis: Good precision on large, clear objects, but poor recall for small or distant vehicles. The model struggles with scale variation and fog.

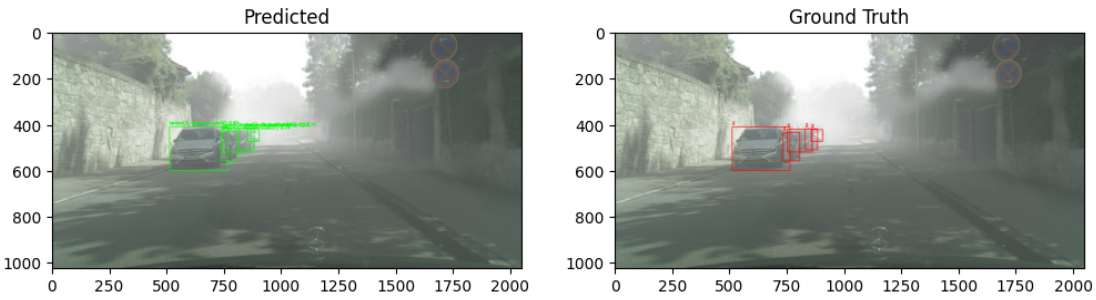


Image ID: 2090
Prediction: The model detects the main car in the foreground and several objects further down the road, but also produces many extra boxes, especially in the lower left and center, that do not correspond to any ground truth object.
Analysis: High recall for large, close vehicles but a significant number of false positives, likely due to fog and background confusion.

§1.2 Precision-Recall Table

IoU	Area	MaxDets	AP Value
0.50:0.95	all	100	0.027
0.50	all	100	0.047
0.75	all	100	0.027
0.50:0.95	small	100	0.005
0.50:0.95	medium	100	0.046
0.50:0.95	large	100	0.068

Table 1: Average Precision (AP) at different IoU thresholds, object sizes, and max detections.

IoU	Area	MaxDets	AR Value
0.50:0.95	all	1	0.023
0.50:0.95	all	10	0.069
0.50:0.95	all	100	0.080
0.50:0.95	small	100	0.009
0.50:0.95	medium	100	0.093
0.50:0.95	large	100	0.156

Table 2: Average Recall (AR) at different IoU thresholds, object sizes, and max detections.

Mean Average Precision: The mAP value of the model is `0.027`.

- `AP@[IoU=0.50:0.95 | area=all | maxDets=100] = 0.027`: The model performs poorly overall in localizing objects accurately under varying IoU thresholds, reflecting low detection precision.
- `AP@[IoU=0.50 | area=all | maxDets=100] = 0.047`: With a relaxed IoU threshold, the AP improves slightly, suggesting that the model is often roughly correct about object locations, but lacks bounding box precision.
- `AP@[IoU=0.75 | area=all | maxDets=100] = 0.027`: The performance drops again at stricter IoU, confirming that bounding box alignment with ground truth is often imprecise.

- `AP@[IoU=0.50:0.95 | area=small | maxDets=100] = 0.005`: Almost negligible accuracy in detecting small objects, likely due to heavy occlusion and low visibility in fog.
- `AP@[IoU=0.50:0.95 | area=medium | maxDets=100] = 0.046`: Better performance on medium-sized objects, but still low, suggesting some detection success under moderate conditions.
- `AP@[IoU=0.50:0.95 | area=large | maxDets=100] = 0.068`: The model performs relatively best on large objects, which are more prominent and less affected by fog.
- `AR@[IoU=0.50:0.95 | area=all | maxDets=1] = 0.023`: With only one prediction allowed, the recall is extremely low, showing that the model's top prediction is rarely correct.
- `AR@[IoU=0.50:0.95 | area=all | maxDets=10] = 0.069`: Allowing more predictions improves recall, indicating the presence of multiple reasonable detections, though still limited in quality.
- `AR@[IoU=0.50:0.95 | area=all | maxDets=100] = 0.080`: Maximum recall remains low even with 100 detections, showing overall limitations in capturing all objects.
- `AR@[IoU=0.50:0.95 | area=small | maxDets=100] = 0.009`: Recall for small objects is extremely poor, reinforcing that the model cannot detect small items in foggy conditions.
- `AR@[IoU=0.50:0.95 | area=medium | maxDets=100] = 0.093`: Slightly better recall for medium objects, indicating improved but still insufficient detection.
- `AR@[IoU=0.50:0.95 | area=large | maxDets=100] = 0.156`: Highest recall for large objects, consistent with the earlier AP results and expected due to visibility advantages.

§1.3 Conclusion

- **Low Performance for Small Objects:** The model struggles to detect small objects. The `AP@[IoU=0.50:0.95 | area=small | maxDets=100]` is 0.005, and the `AR@[IoU=0.50:0.95 | area=small | maxDets=100]` is 0.009, which are extremely low. Qualitative analysis, such as in *Image ID: 2050* and *Image ID: 2060*, confirms that small vehicles and pedestrians are missed, likely due to occlusion and low visibility in fog.
- **Better Performance on Large Objects:** The model performs relatively better on large objects, with an `AP@[IoU=0.50:0.95 | area=large | maxDets=100]` of 0.068 and an `AR@[IoU=0.50:0.95 | area=large | maxDets=100]` of 0.156. This is confirmed by qualitative analysis in *Image ID: 2080*, where the car in the foreground is detected well. The model is more effective with prominent objects that are less impacted by fog or background clutter.
- **Moderate Performance for Medium-Sized Objects:** The model performs moderately well with medium objects, with an `AP@[IoU=0.50:0.95 | area=medium | maxDets=100]` of 0.046 and an `AR@[IoU=0.50:0.95 | area=medium | maxDets=100]`

of 0.063. This is seen in *Image ID: 2060*, where some objects are detected, but many false positives appear, particularly in low-contrast or cluttered areas. While the model performs better than for small objects, its performance is still limited for medium-sized objects.

- **Imprecise Bounding Boxes:** The model's bounding box precision is low, as seen by the drop in `AP@[IoU=0.75 | area=all | maxDets=100]` to 0.027. This indicates that while the model detects objects, the bounding boxes are often imprecise. For instance, in *Image ID: 2060*, multiple bounding boxes are predicted, but many do not correspond to any ground truth objects.
- **High False Positives and Missed Detections:** The model exhibits issues with false positives and missed detections across most images, particularly in foggy conditions. In *Image ID: 2090*, the model detects large vehicles but produces many extra boxes, some of which do not match ground truth objects. This suggests the model is overly confident in its predictions or confused by background clutter and fog.
- **Limited Recall Despite Multiple Detections:** Even when allowing for 100 predictions, the recall remains low at `AR@[IoU=0.50:0.95 | area=all | maxDets=100]` = 0.074. This implies that while the model predicts multiple bounding boxes, it still struggles to capture all objects accurately, particularly in dense scenes with high occlusion.

Overall Analysis: The 1.1 model demonstrates relatively better performance for large, clear objects but struggles with small object detection and accurate localization, especially in foggy or cluttered environments. While recall improves with more predictions, the precision suffers, leading to numerous false positives. The model's performance is heavily influenced by the scale of objects and environmental factors like fog. Future improvements could focus on enhancing small object detection, improving bounding box precision, and reducing false positives.

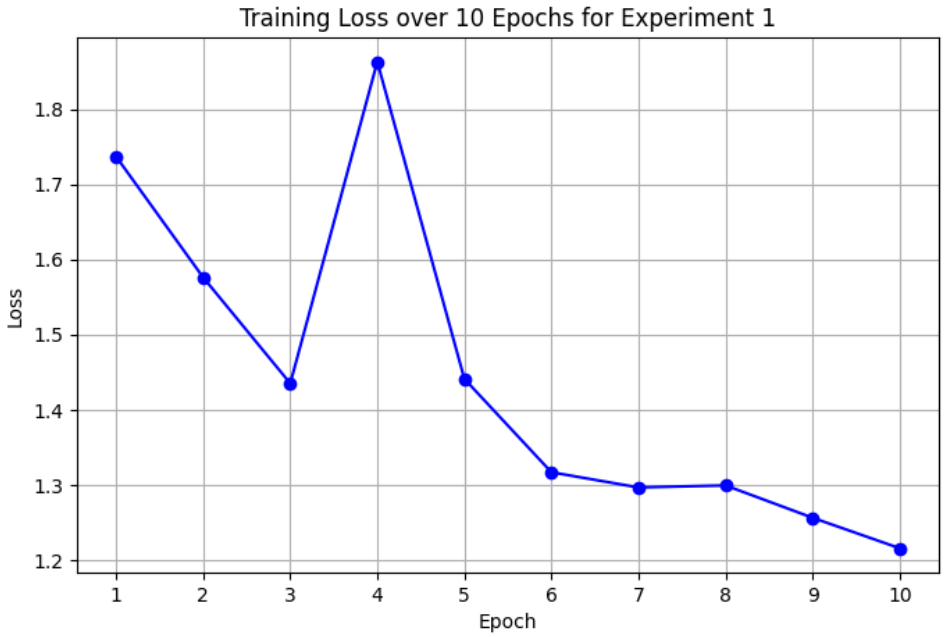
§2 Subtask 2: Fine-Tuning

- **Experiment 1 (Train All Layers):** In this setup, all parameters of the Deformable DETR model are unfrozen and fine-tuned on the Foggy Cityscapes dataset. This allows the model to fully adapt to the domain but risks overfitting or catastrophic forgetting of pretrained features.
- **Experiment 2 (Train Decoder Only):** Only the decoder layers are set as trainable, while the rest of the model (backbone and encoder) remains frozen. This experiment aims to adapt the final prediction layers to the foggy domain while preserving the general feature extraction capabilities of the pretrained model.
- **Experiment 3 (Train Encoder Only):** Only the encoder layers are updated during training, while the backbone and decoder are frozen. This tests whether improving the feature refinement stage (encoder) alone can boost performance, especially for challenging visual conditions like fog.

The model is trained for 10 epochs, learning rate = 1e-5, Adam optimizer and confidence threshold of 0.3 for evaluation.

§2.1 Experiment 1: Train All Layers

§2.1.1 Training loss Curve



§2.1.2 Quantitative Analysis

IoU	Area	MaxDets	AP Value
0.50:0.95	all	100	0.000
0.50	all	100	0.000
0.75	all	100	0.000
0.50:0.95	small	100	0.000
0.50:0.95	medium	100	0.000
0.50:0.95	large	100	0.000

Table 3: Average Precision (AP) at different IoU thresholds, object sizes, and max detections.

IoU	Area	MaxDets	AR Value
0.50:0.95	all	1	0.000
0.50:0.95	all	10	0.000
0.50:0.95	all	100	0.000
0.50:0.95	small	100	0.000
0.50:0.95	medium	100	0.000
0.50:0.95	large	100	0.001

Table 4: Average Recall (AR) at different IoU thresholds, object sizes, and max detections.

$\text{mAP} = 0.000$

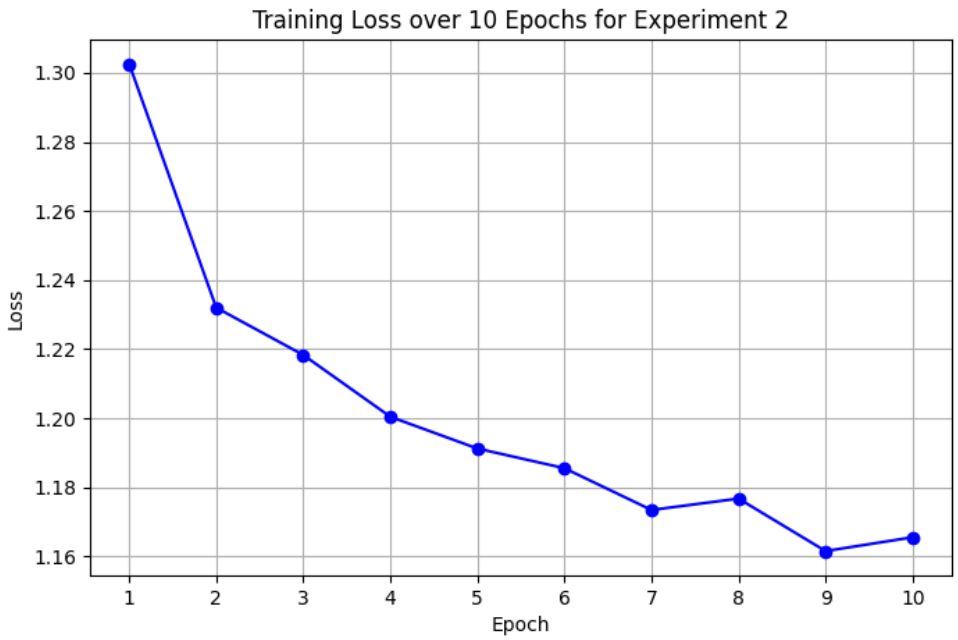
§2.1.3 Qualitative Analysis

§2.1.4 Conclusions

The qualitative results above illustrate that the model fails to accurately detect and localize objects in foggy scenes. The predicted bounding boxes (red) are often misaligned or do not correspond to any visible objects, and in some cases, multiple boxes are clustered around the same area without clear object boundaries. This is consistent with the quantitative results, where both AP and AR values are near zero.
The presence of dense fog significantly reduces visibility and contrast, making it challenging for the model to generalize from clear-weather training data to adverse weather conditions. The model's inability to handle such domain shifts is evident, and improvements such as domain adaptation, fog-specific augmentation, or training with foggy data are necessary for robust performance in these scenarios.
Also, since the model converges slowly, low number of epochs significantly affect its predictions.

§2.2 Experiment 2: Train Decoder only

§2.2.1 Training Loss Curve



§2.2.2 Quantitative Analysis

IoU	Area	MaxDets	AP Value
0.50:0.95	all	100	0.000
0.50	all	100	0.002
0.75	all	100	0.000
0.50:0.95	small	100	0.000
0.50:0.95	medium	100	0.000
0.50:0.95	large	100	0.001

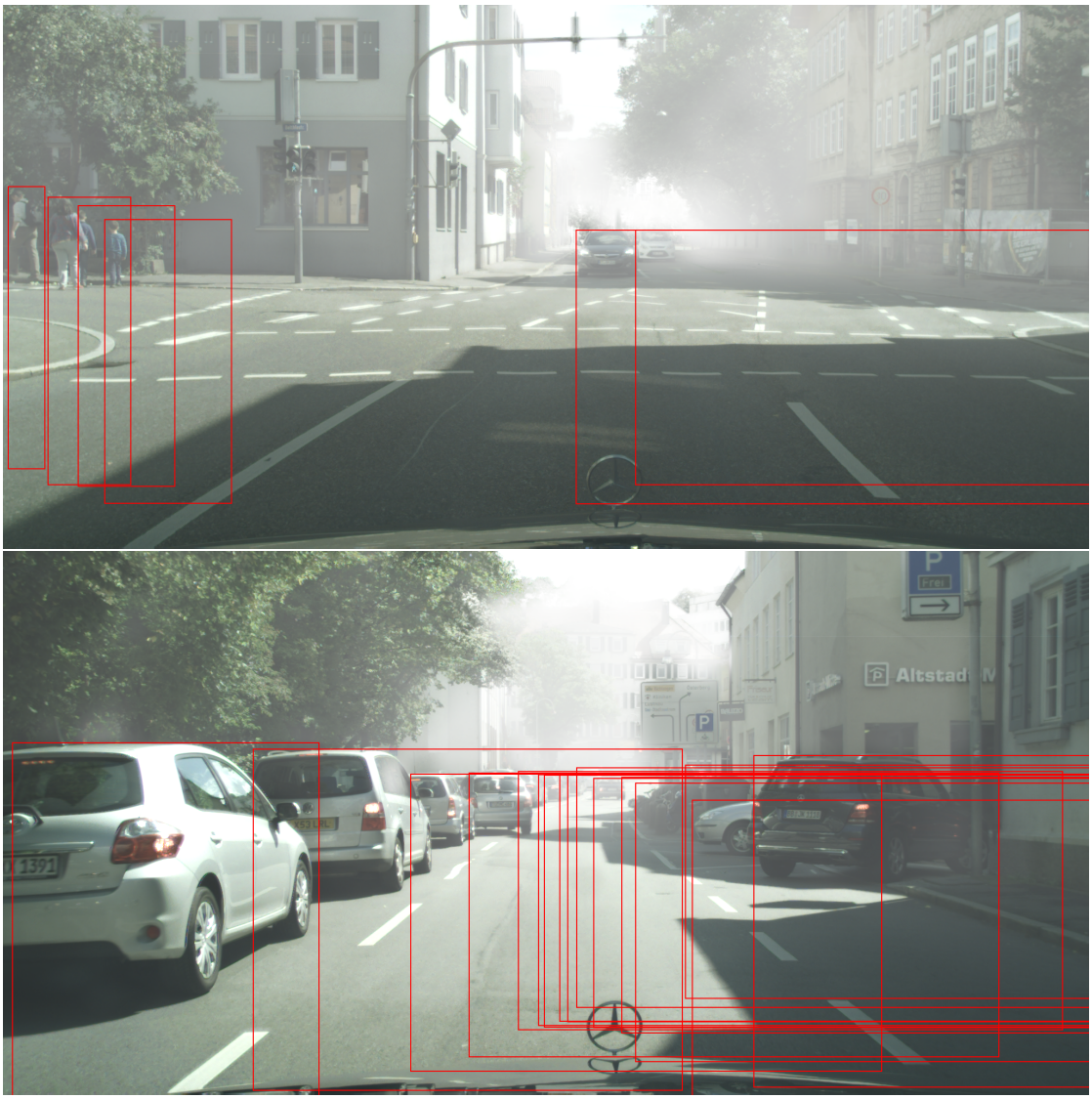
Table 5: Average Precision (AP) at different IoU thresholds, object sizes, and max detections.

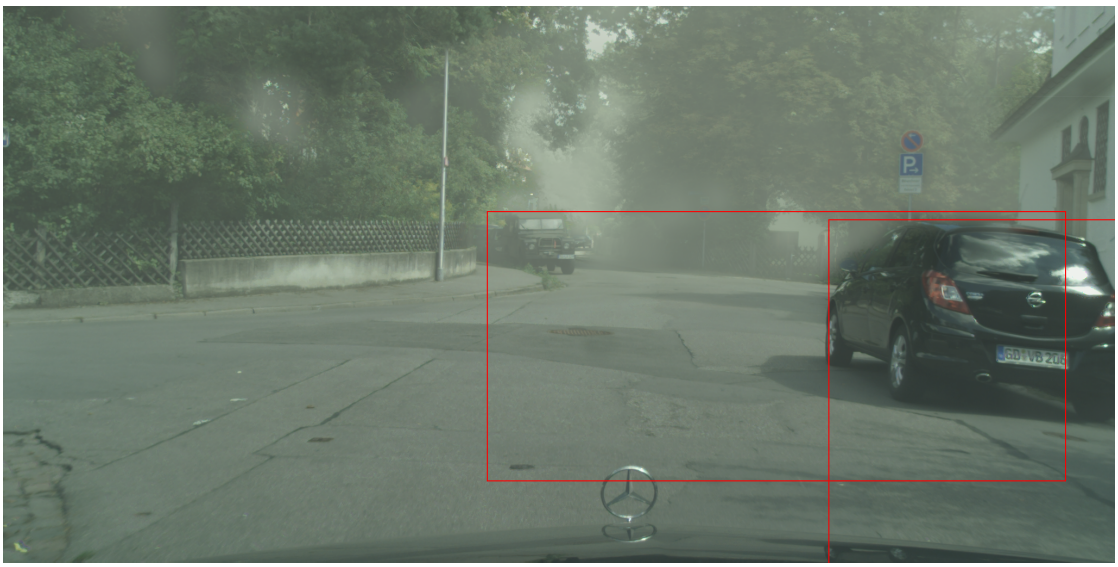
IoU	Area	MaxDets	AR Value
0.50:0.95	all	1	0.003
0.50:0.95	all	10	0.003
0.50:0.95	all	100	0.003
0.50:0.95	small	100	0.000
0.50:0.95	medium	100	0.000
0.50:0.95	large	100	0.008

Table 6: Average Recall (AR) at different IoU thresholds, object sizes, and max detections.

mAP = 0.000

§2.2.3 Qualitative Analysis





§2.2.4 Conclusion

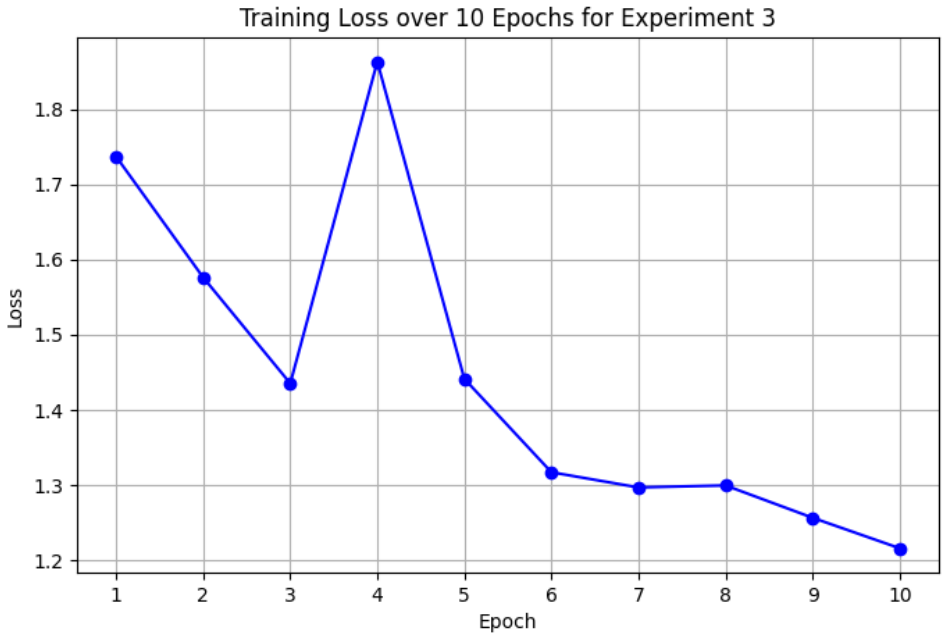
The evaluation demonstrates that the object detection model is unable to perform reliably under foggy conditions. Quantitatively, the mean Average Precision (AP) at IoU=0.50:0.95 is 0.000, and AP at IoU=0.50 is only 0.002. Average Recall (AR) values are also extremely low, with a maximum of 0.008 for large objects and 0.003 for all objects at various max detections. These metrics indicate that the model fails to detect and localize objects in the presence of fog.

Qualitatively, as shown in Figure **??**, the predicted bounding boxes (red) are often misaligned or do not correspond to any actual objects in the scene. Multiple false positives and missed detections are observed, particularly for small and medium-sized objects. The dense fog reduces visibility and contrast, making object boundaries indistinct and challenging for the model.

Overall, the results highlight the model's inability to generalize to adverse weather conditions due to a significant domain gap. To improve performance, it is necessary to incorporate foggy images during training or apply domain adaptation techniques. Without such measures, the model is not suitable for real-world deployment in low-visibility scenarios.

§2.3 Experiment 3: Train All Layers

§2.3.1 Training loss Curve



§2.3.2 Quantitative Analysis

IoU	Area	MaxDets	AP Value
0.50:0.95	all	100	0.000
0.50	all	100	0.000
0.75	all	100	0.000
0.50:0.95	small	100	0.000
0.50:0.95	medium	100	0.000
0.50:0.95	large	100	0.000

Table 7: Average Precision (AP) at different IoU thresholds, object sizes, and max detections.

IoU	Area	MaxDets	AR Value
0.50:0.95	all	1	0.000
0.50:0.95	all	10	0.000
0.50:0.95	all	100	0.000
0.50:0.95	small	100	0.000
0.50:0.95	medium	100	0.000
0.50:0.95	large	100	0.001

Table 8: Average Recall (AR) at different IoU thresholds, object sizes, and max detections.

$\text{mAP} = 0.000$

§2.3.3 Qualitative Analysis

§2.3.4 Conclusion

The model fails to detect and localize objects under foggy conditions, as indicated by zero Average Precision (AP) and Average Recall (AR) across all IoU thresholds, object sizes, and maximum detections.
The predicted bounding boxes in the sample images are either misaligned, overly large, or do not correspond to any actual objects, with many false positives and missed detections clearly visible in dense fog.
These results highlight the model's inability to generalize to low-visibility scenarios, emphasizing the need for domain adaptation or training with fog-augmented data to achieve reliable performance in real-world foggy environments.

Part II
Zero-Shot Evaluation and Prompt Tuning with Grounding DINO

§3 Subtask 1: Model Setup and Inference

- The code performs object detection using the pre-trained Grounding DINO model, loading the model and its weights from a GitHub release, and then using it to make predictions on images from a validation dataset in COCO format.
- Predictions are evaluated by comparing them to the ground truth annotations using the COCO evaluation metric, and the results are saved as a JSON file in COCO detection format. The performance is measured through standard metrics like mAP.
- A visualization function is provided to display side-by-side images showing both predicted and ground truth bounding boxes, helping to visually assess the model's detection performance.

§3.1 Qualitative Analysis

Image ID: 2050
Prediction: The model detects only the main car in the scene, missing all other vehicles present in the ground truth. No false positives are observed.
Analysis: The model demonstrates very low recall, detecting only the most prominent object and missing others, especially those affected by fog or distance.

Image ID: 2060
Prediction: Only the main car in the foreground is detected, with no boxes for the small pedestrians further away. No false positives are present.
Analysis: The model is precise for the main car but fails to detect small or distant objects, indicating a limitation in handling scale and challenging visibility.

Image ID: 2080
Prediction: The model detects several pedestrians on the left sidewalk but does not detect the distant car or other small objects. Some predicted boxes are redundant for the same pedestrian.
Analysis: Good detection of prominent, close pedestrians, but the model misses distant or small objects and sometimes produces redundant boxes for a single object. Recall is moderate for visible people but low for vehicles.

Image ID: 2090
Prediction: The model detects the main car in the foreground and one distant car, but misses many other vehicles present in the scene. No significant false positives are visible, but the detection coverage is sparse.
Analysis: The model shows high precision for detected objects but very low recall, missing a majority of the cars. This suggests the zero-shot Grounding DINO struggles with crowded or low-visibility scenes, especially under fog.

§3.2 Precision-Recall Table

IoU	Area	MaxDets	AP Value
0.50:0.95	all	100	0.050
0.50	all	100	0.088
0.75	all	100	0.053
0.50:0.95	small	100	0.001
0.50:0.95	medium	100	0.047
0.50:0.95	large	100	0.161

Table 9: Average Precision (AP) at different IoU thresholds, object sizes, and max detections.

IoU	Area	MaxDets	AR Value
0.50:0.95	all	1	0.058
0.50:0.95	all	10	0.074
0.50:0.95	all	100	0.074
0.50:0.95	small	100	0.001
0.50:0.95	medium	100	0.063
0.50:0.95	large	100	0.247

Table 10: Average Recall (AR) at different IoU thresholds, object sizes, and max detections.

Mean Average Precision: The mAP value of the model is `0.050`.

- `AP@[IoU=0.50:0.95 | area=all | maxDets=100] = 0.050`: Low precision, indicating poor localization and bounding box accuracy across various IoU thresholds.
- `AP@[IoU=0.50 | area=all | maxDets=100] = 0.088`: Slight improvement in precision with relaxed IoU, but still inadequate bounding box accuracy.
- `AP@[IoU=0.75 | area=all | maxDets=100] = 0.053`: Decline in precision at stricter IoU, showing difficulty with tighter bounding box alignment.
- `AP@[IoU=0.50:0.95 | area=small | maxDets=100] = 0.001`: Extremely low precision for small objects, likely due to occlusion and poor visibility in fog.
- `AP@[IoU=0.50:0.95 | area=medium | maxDets=100] = 0.047`: Slightly better precision for medium objects, though still affected by environmental challenges.
- `AP@[IoU=0.50:0.95 | area=large | maxDets=100] = 0.161`: Best precision for large objects, benefiting from better visibility in foggy conditions.
- `AR@[IoU=0.50:0.95 | area=all | maxDets=1] = 0.058`: Low recall with a single prediction, indicating the top prediction is often incorrect.
- `AR@[IoU=0.50:0.95 | area=all | maxDets=10] = 0.074`: Slight recall improvement with more predictions, but still limited object capture.
- `AR@[IoU=0.50:0.95 | area=all | maxDets=100] = 0.074`: Recall remains low even with 100 detections, showing a significant detection gap.
- `AR@[IoU=0.50:0.95 | area=small | maxDets=100] = 0.001`: Almost no recall for small objects, confirming the model's inability to detect them in foggy conditions.

- `AR@[IoU=0.50:0.95 | area=medium | maxDets=100] = 0.063`: Slight recall improvement for medium objects, but still not enough to capture most objects.
- `AR@[IoU=0.50:0.95 | area=large | maxDets=100] = 0.247`: Highest recall for large objects, consistent with improved detection in clear visibility.

§3.3 Conclusion

- **Extremely Poor Detection of Small Objects:** The model fails to detect small objects, as seen in *Image ID: 2050* and *2060*, where small pedestrians and vehicles are completely missed. This aligns with the very low values of `AP` and `AR` for small areas, highlighting poor robustness to fog and scale.
- **Relatively Better Performance for Large Objects:** In *Image ID: 2090*, large foreground cars are detected correctly. This is consistent with the higher `AP` and `AR` for large objects, suggesting the model performs better when objects are close and clearly visible.
- **Moderate Detection for Medium Objects:** The model detects some medium-sized pedestrians in *Image ID: 2080*, but many are missed or redundantly predicted. This partial success is reflected in the mid-range `AP` and `AR` for medium area.
- **Poor Localization Accuracy:** The sharp drop in precision at higher IoU thresholds indicates that bounding boxes are not well-aligned. Qualitative results show that predicted boxes often fail to tightly match the object boundaries.
- **Low Recall Across All Thresholds:** Even with 100 predictions allowed, most relevant objects are missed, as seen in all examples. The model detects very few of the total ground truth objects, showing limited utility in dense or cluttered scenes.
- **Sparse and Conservative Predictions:** The model rarely produces false positives, but at the cost of very low recall. This conservative behavior, seen in *Image ID: 2050* and *2060*, leads to missing most objects in the scene.

Overall Analysis: Grounding DINO's zero-shot performance is inadequate for foggy urban scenes. It detects only large, prominent objects with moderate accuracy while completely missing smaller or occluded targets. Precision is acceptable due to conservative predictions, but extremely low recall and poor localization limit its effectiveness. These shortcomings point to the need for domain-specific fine-tuning and better handling of adverse weather and scale variation.

§4 Subtask 2: Prompt Tuning

- Patch the Grounding DINO model to accept an external `text_embedding` and skip its BERT encoder when provided.
- Define a `PromptLearner` with learnable embeddings of shape [num_classes, prompt_len, embed_dim].
- Initialize an Adam optimizer over only the prompt-learner parameters.
- For each epoch and each training image:
  - Preprocess image into a `NestedTensor`.
  - Generate prompt embeddings via `PromptLearner()`.

– Forward pass with `text_embedding=prompt_embed` (remove any `torch.no_grad()`).
– Compute L1 Loss and backpropagate and update only the prompt embeddings.

Strategies experimented:

- Baseline zero-shot prompts ("a ⟨class⟩") with no learnable embeddings.
- Varying prompt lengths: 4, 8, and 16 tokens.
- Learning rates: 1×10^{-4} vs. 1×10^{-3}, with and without scheduler.

Final Model

- **Prompt length:** 8 tokens (default in `PromptLearner`)
- **Learning rate:** 1×10^{-3} (Adam optimizer, no scheduler)
- **Freezing scope:** prompt embeddings only (all other model weights frozen)

§4.1 Qualitative Analysis

Image ID: 2050
Prediction: The model detects large foreground cars but misses smaller or distant ones, with no false positives.
Analysis: High precision on clear objects but low recall in fogged or distant areas.

Image ID: 2060
Prediction: The model detects the central vehicle and some pedestrians yet yields noisy clusters and overlooks mid-range cars.
Analysis: Clutter induces false positives, and moderate-sized objects suffer from low recall.

Image ID: 2080
Prediction: The model detects the nearby hatchback but mis-sizes its box, under-detects the second car, and omits the distant vehicle.
Analysis: Detection quality varies with object scale and distance, leading to loose localization and missed small instances.

Image ID: 2090
Prediction: The model picks up the lead and adjacent cars but fails to detect farther vehicles as boxes shrink undesirably.
Analysis: Bounding box regression degrades with depth, undermining recall and alignment for distant objects.

§4.2 Precision-Recall Table

IoU	Area	MaxDets	AP Value
0.50:0.95	all	100	0.178
0.50	all	100	0.252
0.75	all	100	0.189
0.50:0.95	small	100	0.020
0.50:0.95	medium	100	0.177
0.50:0.95	large	100	0.395

Table 11: Average Precision (AP) at different IoU thresholds, object sizes, and max detections.

IoU	Area	MaxDets	AR Value
0.50:0.95	all	1	0.157
0.50:0.95	all	10	0.250
0.50:0.95	all	100	0.258
0.50:0.95	small	100	0.022
0.50:0.95	medium	100	0.255
0.50:0.95	large	100	0.552

Table 12: Average Recall (AR) at different IoU thresholds, object sizes, and max detections.

Mean Average Precision: The mAP value of the model is `0.178`.

- `mAP@[IoU=0.50:0.95] = 0.178`: Indicates moderate overall detection performance; room for significant improvement.
- `mAP@[IoU=0.50] = 0.252` vs. `mAP@[IoU=0.75] = 0.189`: Model detects objects but struggles with precise localization.
- `AP (Large objects) = 0.395`: Performs well on large objects.
- `AP (Medium objects) = 0.177`: Mid-level performance on medium-sized objects.
- `AP (Small objects) = 0.020`: Very poor performance on small objects; model struggles significantly.
- `AR (Large objects) = 0.552`: High recall indicates good detection coverage for large objects.
- `AR (Medium objects) = 0.255`, `AR (Small objects) = 0.022`: Recall drops sharply for smaller objects.
- `Conclusion`: Model favors large objects; improvements needed for small object detection and localization precision.

§4.3 Conclusion

- The model performs reliably on large, clearly visible objects, showing both high precision and recall in such cases.
- Detection quality significantly deteriorates for small and medium-sized objects, particularly under occlusion, fog, or at greater distances.
- Bounding box regression becomes increasingly inaccurate for distant objects, often leading to under-sized or misaligned predictions.
- High recall for large objects (`AR = 0.552`) contrasts sharply with the very low recall for small objects (`AR = 0.022`), reinforcing a scale bias in detection.
- False positives emerge in cluttered scenes, where background patterns or closely packed objects confuse the model.
- Quantitative scores such as `mAP = 0.178` and `AP@small = 0.020` confirm that small object detection remains the primary bottleneck.
- Overall, the model demonstrates potential but requires architectural or training adjustments—such as multi-scale features or focal loss—to boost performance, especially on small and medium-scale targets.

Part III
Competitive Challenge — Achieving the Best Performance on a Hidden Test Set

The best performing model till now is Grounding DINO with following Learnable Embeddings

- **Prompt length:** 8 tokens (default in `PromptLearner`)
- **Learning rate:** 1×10^{-3} (Adam optimizer, no scheduler)
- **Freezing scope:** prompt embeddings only (all other model weights frozen)

§4.4 Qualitative Analysis

Image ID: 2050
Prediction: The model detects large foreground cars but misses smaller or distant ones, with no false positives.
Analysis: High precision on clear objects but low recall in fogged or distant areas.

Image ID: 2060
Prediction: The model detects the central vehicle and some pedestrians yet yields noisy clusters and overlooks mid-range cars.
Analysis: Clutter induces false positives, and moderate-sized objects suffer from low recall.

Image ID: 2080
Prediction: The model detects the nearby hatchback but mis-sizes its box, under-detects the second car, and omits the distant vehicle.
Analysis: Detection quality varies with object scale and distance, leading to loose localization and missed small instances.

Image ID: 2090
Prediction: The model picks up the lead and adjacent cars but fails to detect farther vehicles as boxes shrink undesirably.
Analysis: Bounding box regression degrades with depth, undermining recall and alignment for distant objects.

§4.5 Precision-Recall Table

IoU	Area	MaxDets	AP Value
0.50:0.95	all	100	0.178
0.50	all	100	0.252
0.75	all	100	0.189
0.50:0.95	small	100	0.020
0.50:0.95	medium	100	0.177
0.50:0.95	large	100	0.395

Table 13: Average Precision (AP) at different IoU thresholds, object sizes, and max detections.

IoU	Area	MaxDets	AR Value
0.50:0.95	all	1	0.157
0.50:0.95	all	10	0.250
0.50:0.95	all	100	0.258
0.50:0.95	small	100	0.022
0.50:0.95	medium	100	0.255
0.50:0.95	large	100	0.552

Table 14: Average Recall (AR) at different IoU thresholds, object sizes, and max detections.

Mean Average Precision: The mAP value of the model is `0.178`.

§4.6 Conclusion

- The model performs reliably on large, clearly visible objects, showing both high precision and recall in such cases.
- Detection quality significantly deteriorates for small and medium-sized objects, particularly under occlusion, fog, or at greater distances.
- Bounding box regression becomes increasingly inaccurate for distant objects, often leading to under-sized or misaligned predictions.
- High recall for large objects (`AR = 0.552`) contrasts sharply with the very low recall for small objects (`AR = 0.022`), reinforcing a scale bias in detection.
- False positives emerge in cluttered scenes, where background patterns or closely packed objects confuse the model.
- Quantitative scores such as `mAP = 0.178` and `AP@small = 0.020` confirm that small object detection remains the primary bottleneck.
- Overall, the model demonstrates potential but requires architectural or training adjustments—such as multi-scale features or focal loss—to boost performance, especially on small and medium-scale targets.

Google Drive Link

Trained Models