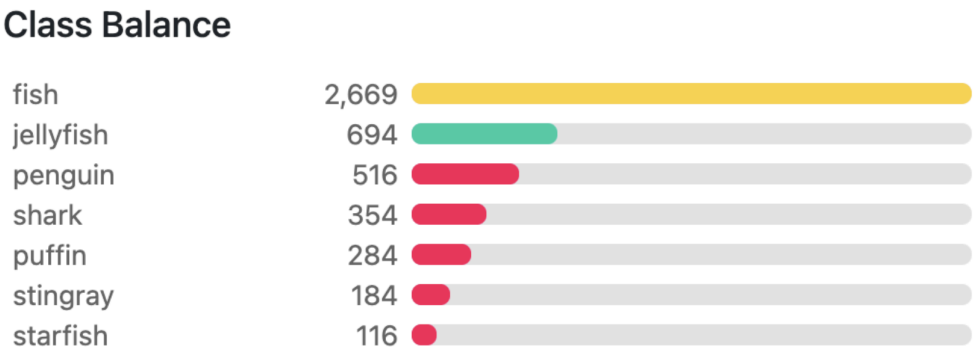
Query-Based Object Detection – Testing & Evaluation

Objective

The goal of this task was to implement a **minimal query-based object detection model** inspired by RT-DETR. The design included a **ResNet-18 backbone** for feature extraction, a **single transformer decoder block** with 12 queries, and a simple prediction head for bounding box regression and classification. The model was trained for 10 epochs on a Roboflow dataset of underwater animals (fish, sharks, starfish).

Dataset

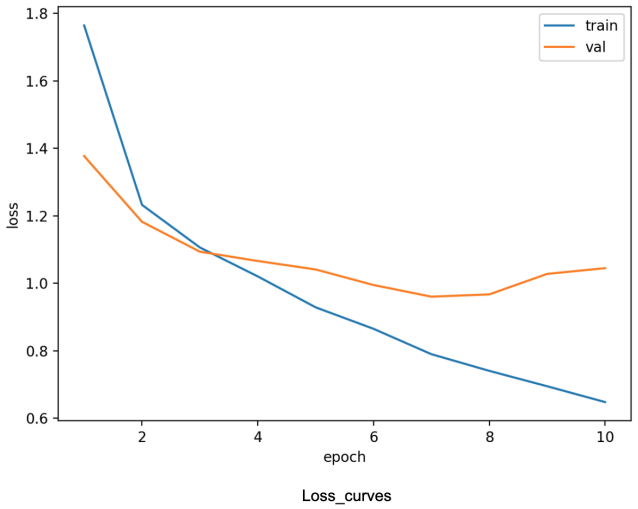
Name: Aquarium Computer Vision Dataset
Link: https://public.roboflow.com/object-detection/aquarium
Images: 638
Classes (7) : Fish, jellyfish, penguin, puffin, shark, starfish, stingray



Experimental Setup (Custom Settings)

- **Backbone:** ResNet-18 pretrained.
- **Decoder:** 1 Transformer decoder block with 12 queries.
- **Training:** 10 epochs, batch size = 16, learning rate = 1e-4.
- **Losses:** CrossEntropy for classification, L1 for bounding boxes.

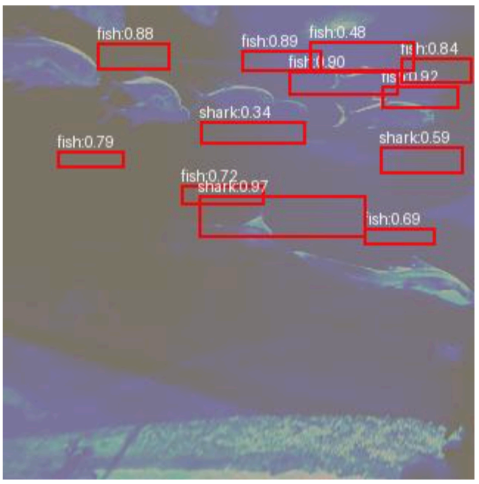
Results

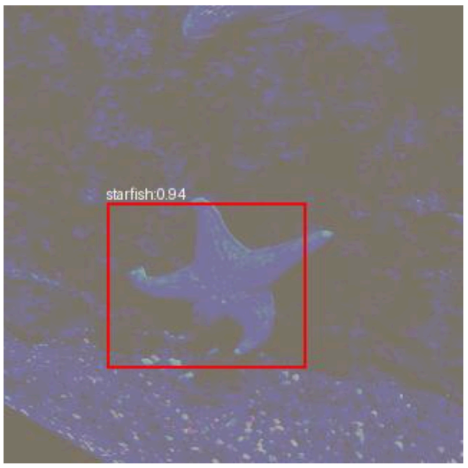


Loss_curves

Loss Curves:loss steadily decreased and validation loss followed but leveled off after epoch 6–7, showing that the model learned patterns but began overfitting slightly.

Sample Detections:





Starfish Detection

- **Starfish Detection (Accurate):**

- The model identified a starfish with **94% probability**, bounding box closely matching the object. This shows strong learning on distinct shapes.
- **Crowded Scene (Partial Accuracy):**
  In multi-object images with fish and sharks, the model produced several bounding boxes, some correct but others overlapping or misplaced. This comes from limited queries (12) and small dataset size, which restrict generalization.

Conclusion

The project successfully delivered a **query-based object detector** matching all assignment requirements. It can detect clear single objects well, but crowded underwater scenes challenge it, leading to overlapping predictions and reduced confidence.

Reflection

This work demonstrates how query-based detection works even in a simplified form. The approach was effective for obvious objects but limited by:

1. Small dataset size.
2. Short training (10 epochs).
3. Minimal model complexity (1 decoder layer).