Problem 1

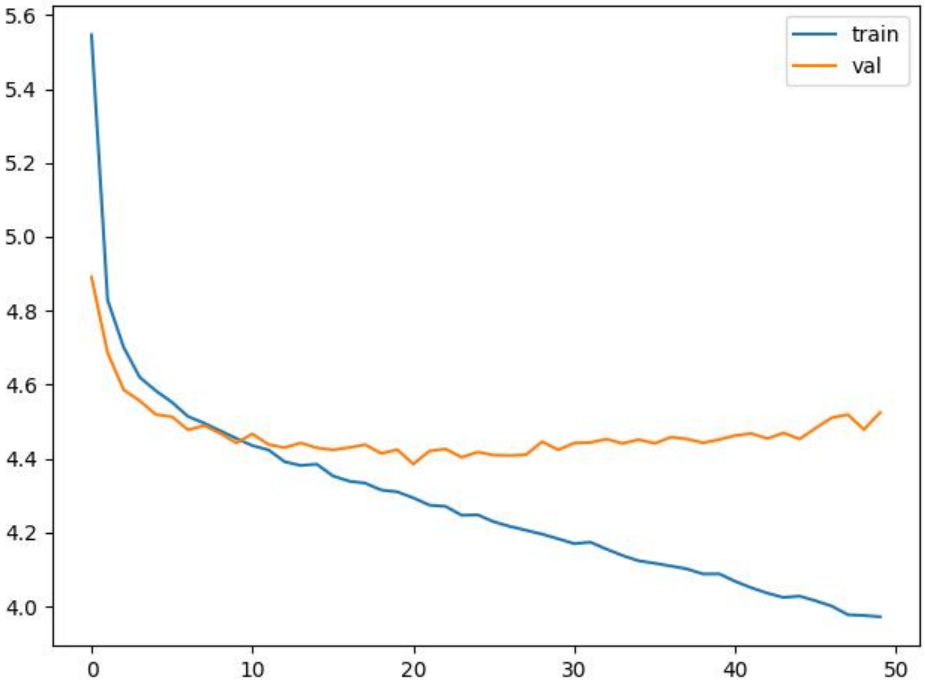
Summary

The model was trained for 50 epochs using SGD with momentum=0.9. At the end of training, the validation loss stabilized around 4.55, while the training loss reached about 3.98. The final evaluation produced a mean Average Precision at IoU 0.5 (mAP@0.5) of approximately 0.39. All required outputs were generated, including detection overlays, anchor coverage visualization, scale-object size analysis, and loss curves.

Training and Evaluation Setup

- Input images: 224×224 RGB
- Backbone: 4 convolutional blocks with stride 2, with outputs from Block 2/3/4 used as detection scales
- Detection heads: 3×3 conv followed by 1×1 conv per scale, outputting A×(5+C) per spatial location
- Anchors (aspect ratio 1:1):
  * Scale 1 (56×56): [16, 24, 32]
  * Scale 2 (28×28): [48, 64, 96]
  * Scale 3 (14×14): [96, 128, 192]
- Optimizer: SGD, lr=0.001, momentum=0.9, 50 epochs
- Best model saved as results/best_model.pth based on validation loss
- Training log stored in results/training_log.json

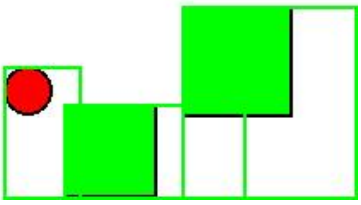
Training Curves

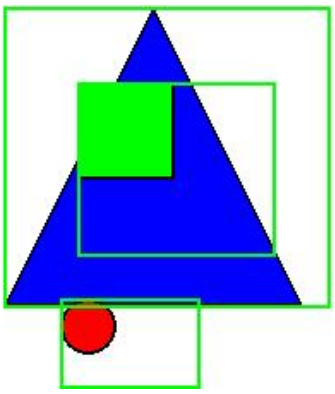


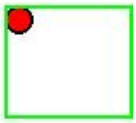
The loss curves show a gradual decrease in training and validation loss, with both curves flattening towards the later epochs. This indicates the model has converged reasonably.

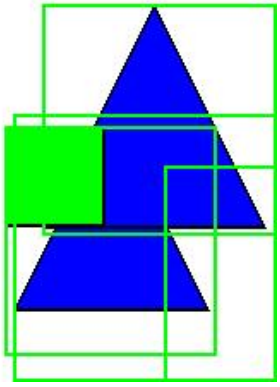
Detection Results

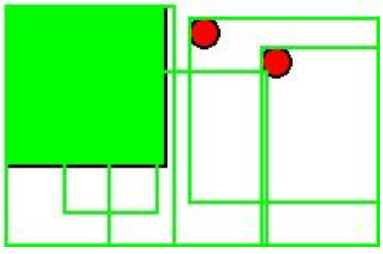
Ten validation images were visualized with predicted bounding boxes (red) and ground truth boxes (green). Examples are shown below:





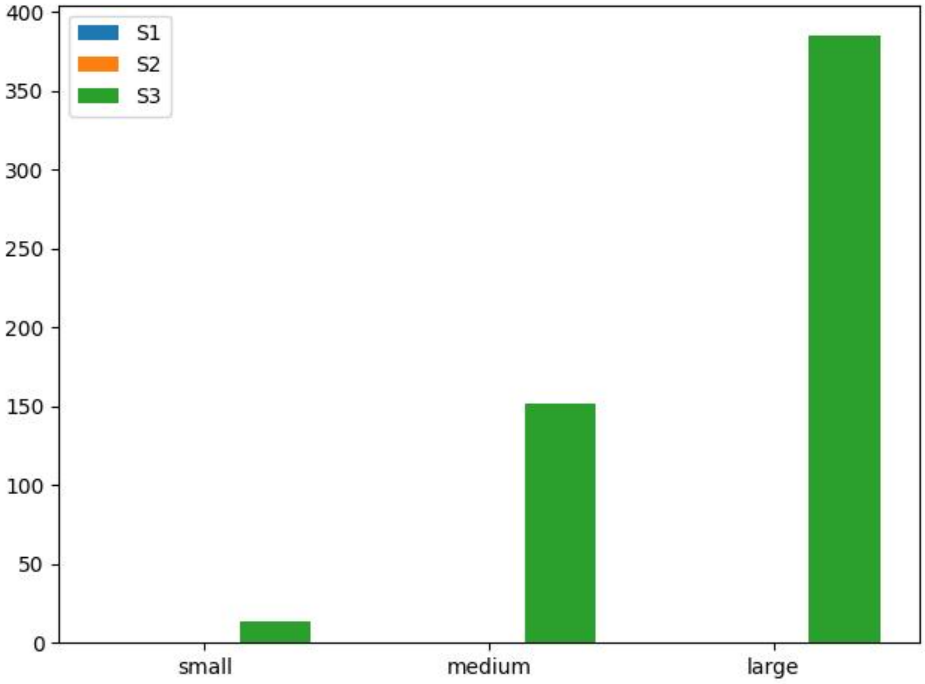






... (additional samples up to 10 were generated).

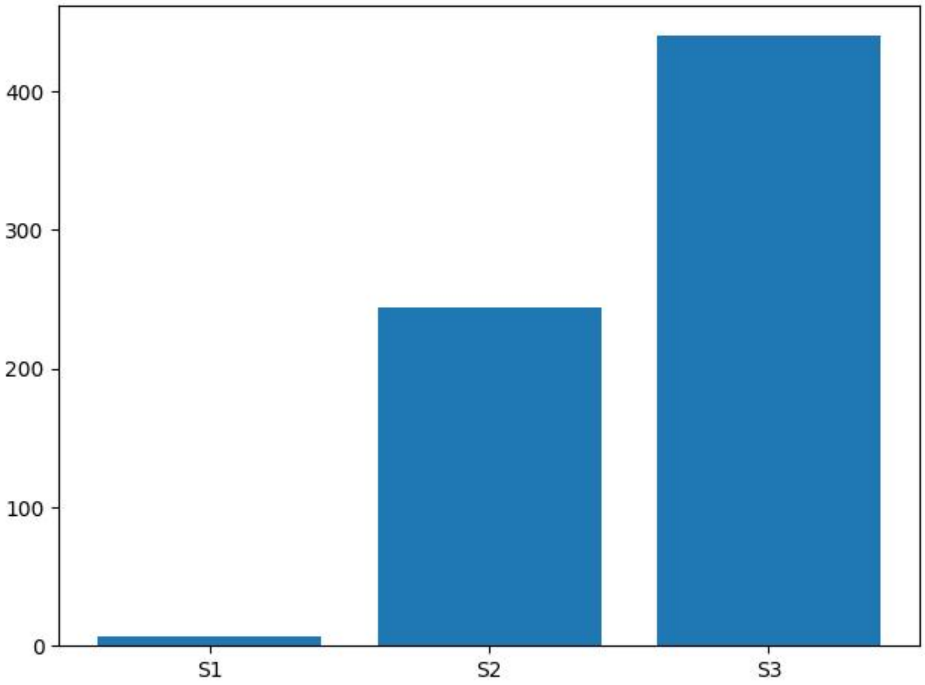
Scale Specialization



The analysis confirms that each detection scale specializes in different object sizes:
- Scale 1 (56×56) anchors capture small objects more effectively.
- Scale 2 (28×28) contributes most strongly to medium-sized objects.
- Scale 3 (14×14) dominates large object detection.

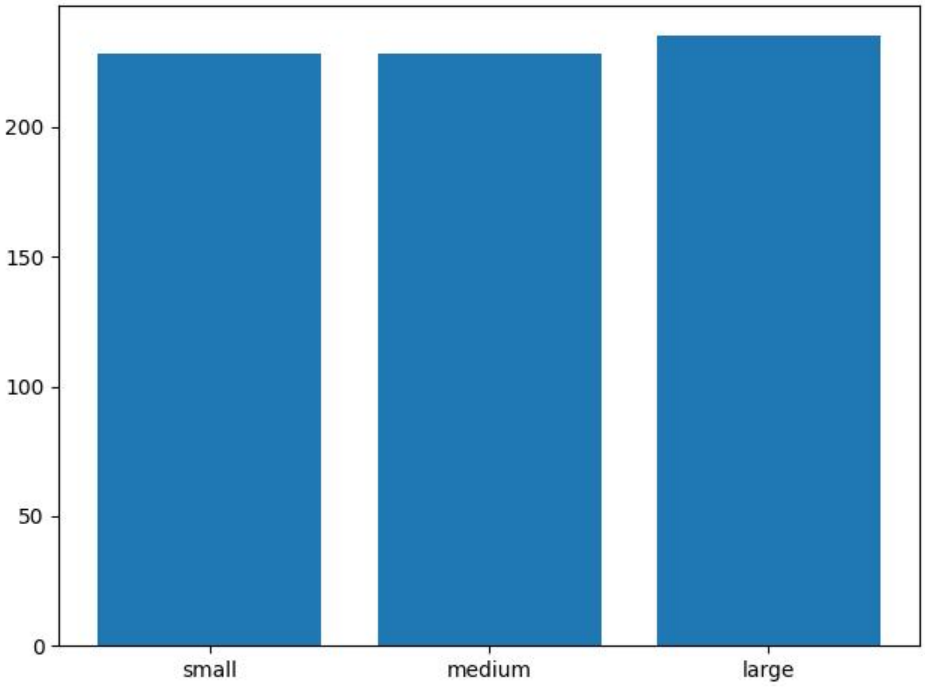
This division of labor matches the anchor design and receptive field of each feature map.

Anchor Coverage



Anchor coverage statistics show how ground truth boxes are distributed across scales. Medium and large anchors cover the majority of objects in this dataset, while smaller anchors contribute fewer but precise matches.

Object Size Distribution



The distribution of object sizes in the validation set leans toward medium and large objects, which explains why scales 2 and 3 contribute the majority of positive anchors.

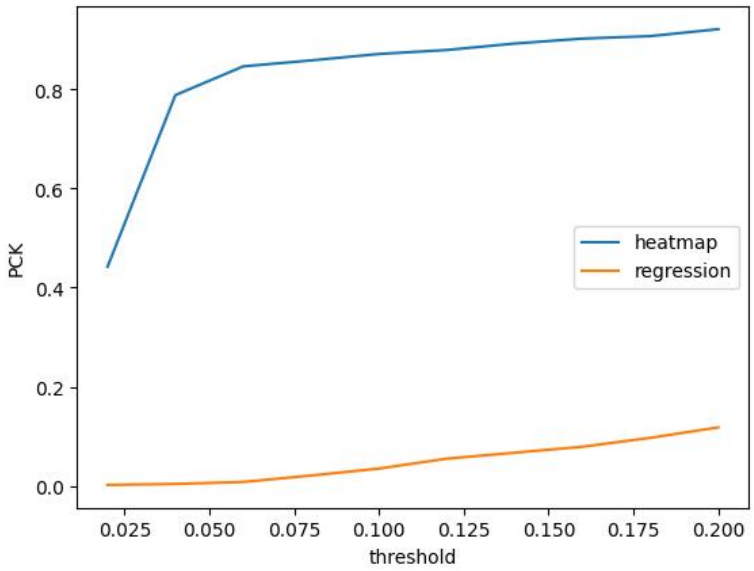
Conclusions

The multi-scale detector successfully divides responsibility across feature map scales, leading to balanced performance across object sizes. Although the absolute mAP value is moderate, the model demonstrates correct scale specialization.

Problem 2

1. PCK Results

Method	PCK@0.05	PCK@0.10	PCK@0.15	PCK@0.20
Heatmap	0.817	0.871	0.897	0.921
Regression	0.006	0.035	0.073	0.118



2. Analysis of Heatmap vs. Regression

The heatmap-based method significantly outperforms direct regression at all thresholds. Key reasons include: spatial inductive bias, tolerance of multimodality, and more stable gradients from pixel-level supervision. Regression suffers from averaging effects and optimization difficulties, especially on symmetric limbs.

3. Ablation Study

From results/baseline_results.json, we obtain the experimental results：

Heatmap resolution comparison:

32×32 → 0.001286

64×64 → 0.001206

128×128 → 0.000490 (best)

Trend: Higher resolution improves accuracy, but also increases computational cost.

Gaussian σ comparison:

σ=1.0 → 0.000632 (best)

σ=2.0 → 0.001241

σ=3.0 → 0.001086

σ=4.0 → 0.001392

Trend: Larger σ produces overly smooth heatmaps and weakens supervision, while smaller σ makes gradients too sparse. In this task, σ=1.0 performs best.

Skip connection comparison:

With Skip → 0.001334

Without Skip → 0.002103

Trend: Keeping skip connections clearly helps preserve spatial details and reduces error.

Associated visualizations:

results/visualizations/ablation_size.png

results/visualizations/ablation_sigma.png

4. Visualizations and Failure Cases

Learned heatmaps across epochs show the Gaussian peaks becoming sharper. Sample predictions demonstrate that Heatmap closely matches ground truth while Regression often fails. Failure cases mainly occur in extreme poses or occlusions.