Problem 1 Analysis:
Specialization of Anchor Scales for Different Object Sizes

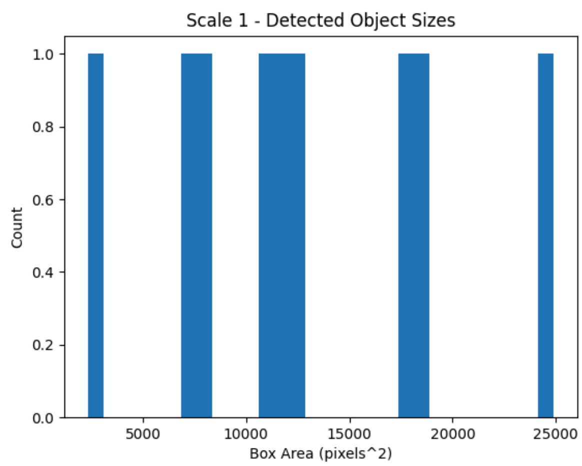

Figure 1

- **Small anchors (Scale 1):**
  These primarily detect objects below 25,000 px², which corresponds to small isolated shapes. These anchors trade precision for the inability to identify larger objects.

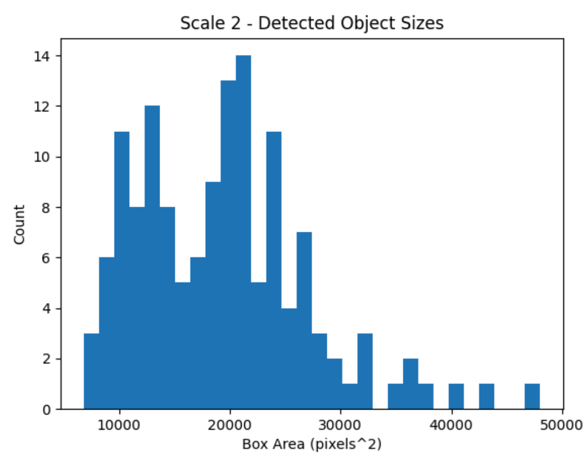

Figure 2

- **Medium anchors (Scale 2):**
  Most effective in the 10,000–30,000 px² range. These anchors provide a balance between localization and receptive field coverage, making them well-suited for medium-sized shapes.

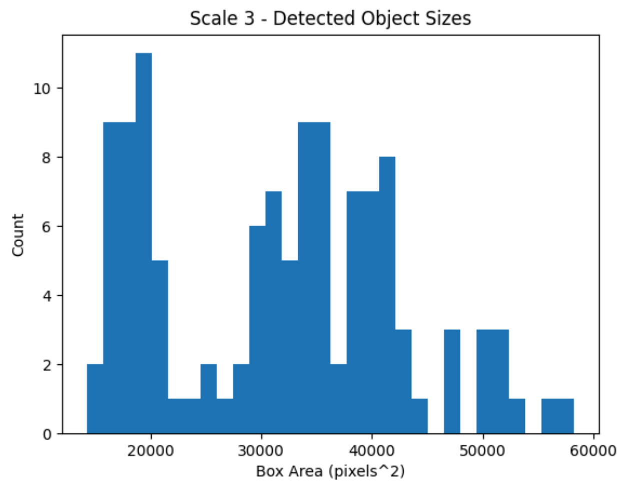

Figure 3

- **Large anchors (Scale 3):**
  Capable of detecting a wide span of objects, from 10,000–60,000 px². Their broader receptive field allows them to capture large structures at the expense of precision on small shapes. Because large anchors sometimes over-predict or overlap multiple smaller objects, smaller anchors are needed.

The Effect of Anchor Scales on Detection Performance

For different types and sizes of shapes, the choices of different anchor scales can matter drastically. For example, in Figure 1, there are very few detections, showing that small anchors cannot identify large objects. Alternatively, even though large anchors can identify small objects as shown in Figure 3, they tend to misalign frequently, and also contain other objects that might be too close.

Interestingly, training logs show the model detects triangles far better than other shapes. This suggests the current anchor scales align more naturally with triangular bounding box distributions, whereas circles and squares may require finer anchor tuning.

Visualization of the Learned Features at Each Scale

Below are two visualizations of validation detections and ground truth boxes that represent the scope of detections very well.

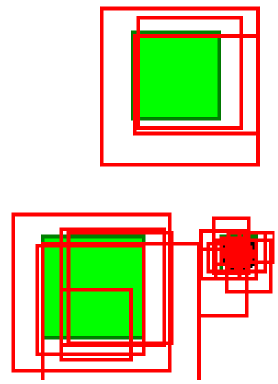

Figure 4

In Figure 4, there are 2 medium-sized shapes and a small one. This confirms that small anchors are more suited for small shapes, while medium/large anchors dominate on larger ones.

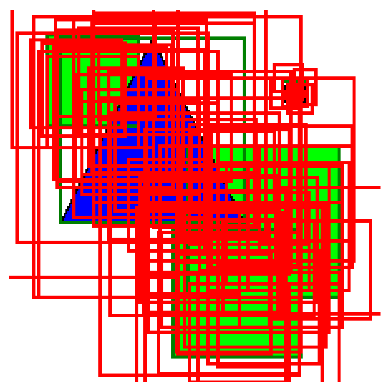

Figure 5

Figure 5 is a lot more difficult for the model to identify, as there are multiple overlapping shapes, therefore demonstrating scale limitations in cluttered contexts. Small anchors succeed on isolated small shapes, but medium and large anchors struggle with object overlap, causing duplicate detections or missed boxes to be false positives.

Problem 2 Analysis:
PCK Curve at Thresholds [0.05, 0.1, 0.15, 0.2]

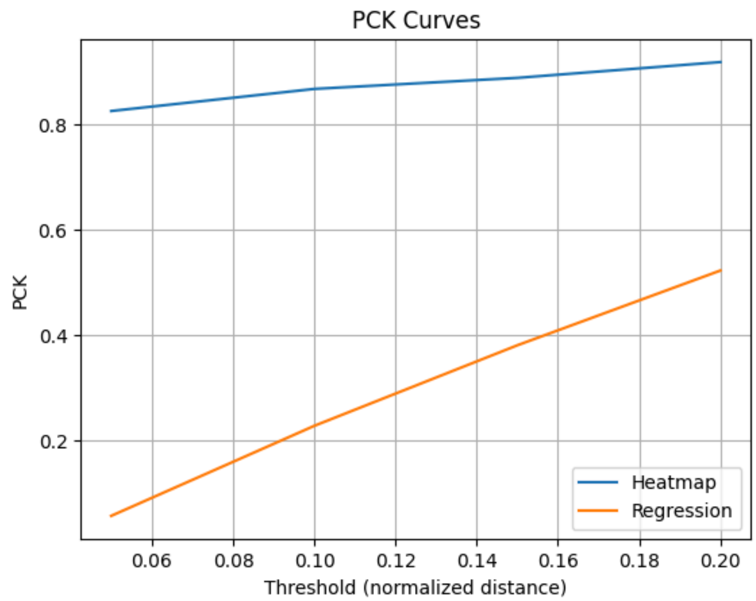

Figure 6

In Figure 6,the heatmap model consistently achieves >80% PCK even at 0.05, and gradually improves with larger thresholds, reaching above 90% by 0.2. Whereas the regression model starts much lower, 5% PCK at 0.05, but shows steady improvement with threshold, reaching 50% PCK at 0.2.

Analysis on Heatmap vs. Regression
From both the PCK curve and the training log, it is easy to see that the heatmap model performs significantly better. It is capable of detecting more keypoints and makes fewer errors. On the other hand, the regression model does have the advantage of having a simpler architecture, making it more efficient to train. However, it cannot reach a high accuracy, and it is also more difficult for it to align predictions.

Ablation Study Results of Sigma, Resolution, and Skip Connections
The resolution study shows that the resolution of the heatmaps matters significantly, that a 32x32 heatmap has double the validation loss in comparison to a 128x128 heatmap. So a more detailed heatmap is generally better for performance.
The sigma study shows that a small sigma achieves the best result, but increasing the sigma value further seems not to affect the validation loss.
The skip connection study shows skipping connections helps the model better identify keypoints, but the difference in validation loss is not very large.

Conclusion: For the best results, use a detailed heapmap(128x128), small sigma(0.1), and skip connections.

Visualization of Learned Heatmaps

The following figures visualize the heatmaps that focus on head detection at epochs 1, 10, and 20:

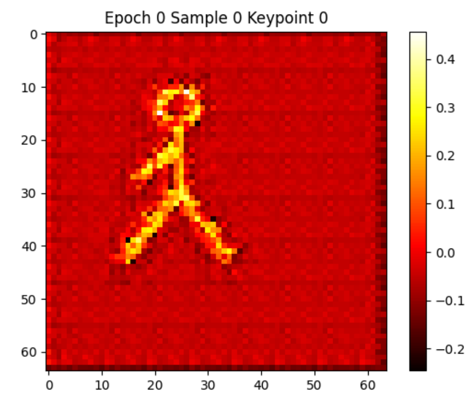

Figure 7

For epoch 0, the model has not learned much, so all features are still present on the heatmap, and there is little to no keypoint separations.

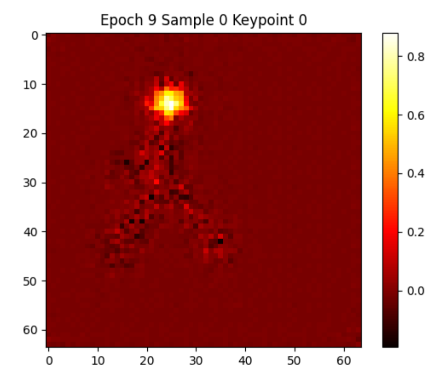

Figure 8

For epoch 10, it is easy to see that the model has already successfully identified the correct keypoint, and the rest of the features are already dulled in comparison, separating the other keypoints from the correct one.

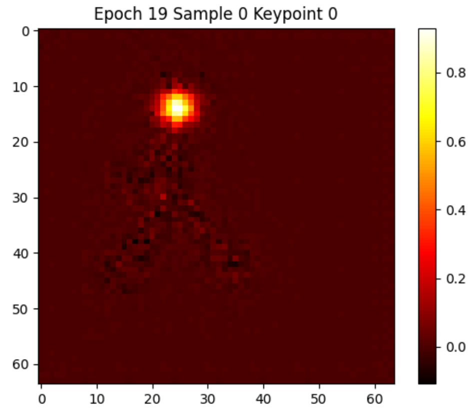

Figure 9

For epoch 20, the heatmap basically only shows a bright correct keypoint, which shows the model has localized the peak and is confident in its detection.

Visualization of Failure Cases

For the failure cases, it only captured the cases where the regression model fails but the heatmap model is good, which makes sense considering the final validation loss of the heatmap model is 0.008. Figure 10 shows one case of regression model failure.

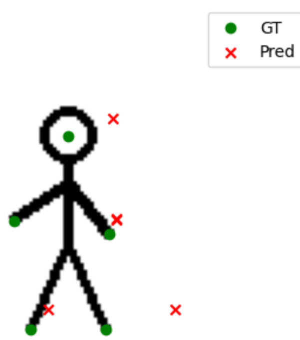

Figure 10

The predictions look shifted to the right, and it also fails to identify the right-hand keypoint, confusing it with the left-hand. Most of the other failure cases are also similar. However, the right shift is also observed in some other validation visualizations, such as Figure 11.

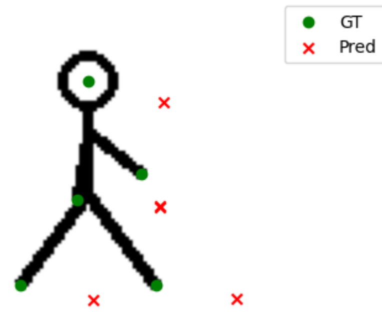

Figure 11

Since the heatmap model has a very high accuracy, it is within reason to believe that right shift might be a problem with the visualization itself. Unfortunately, the reason behind this was not identified.