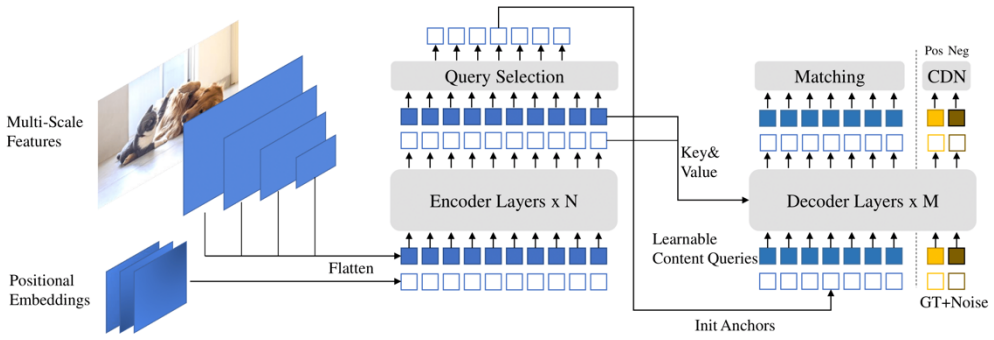
# CVPDL-HW1 ( R12944047 邱彥慈 )

## ● Architecture : DINO

- ○ A DERT-like model composed of backbone + transformer encoders & decoders
  - ▪ From previous models
    - □ Backbone architecture
      - ■ Multiscale features from ResNet, Swin transformer,…
    - □ Deformable attention
    - □ Positional query:
      - ■ 4D anchor box dynamically update through each decoder layer
  - ▪ Innovation
    - □ Mixed Query Selection
      - ■ Select top-K encoder features in the last layer to initialize the positional queries, while the content queries are kept learnable as before.
    - □ Look forward twice
      - ■ In look forward once, the predicted box $b_i^{(pred)} = f(b_{i-1}, \Delta b_i)$, while in look forward twice, $\Delta b_i$ is used to update the box twice, i.e., $b_i'$ and $b_{i+1}^{(pred)}$.
    - □ CDN
      - ■ Contrastive Denoising box with both positive and negative samples.



# Implementation

## Parameter Setting:

- Start from pretrained weights
  - DINO-4scale provides 3 kinds of pretrained wights, which have been trained on COCO for 12, 24, 36 epochs respectively.
    I take the 36-epochs one, which is actually checkpoint0029.pth

- Parameters settings are same as the default, except for

| learning_rate | random_seed | batch_size | epochs |
|---|---|---|---|
| 1e-5 | 0 | 1 | 15 |

  - I've trained 24 epochs on ResNet-DINO and found there's a performance gap at about epoch 12, thus I prefer training to be more than 12 epochs, and it turns out that Swin-DINO takes around 10 epochs to converge.

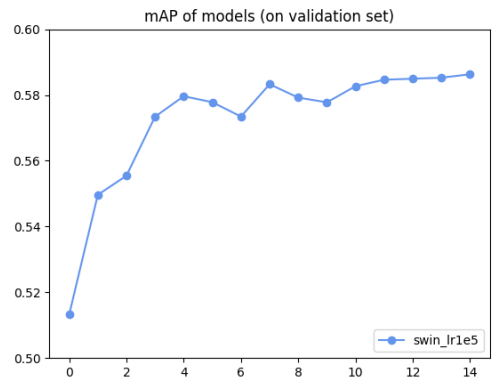
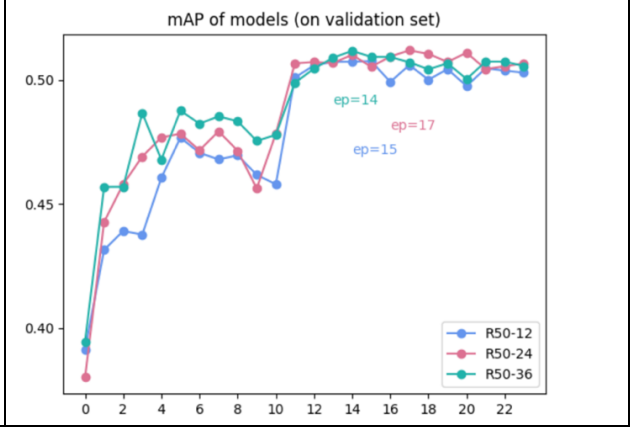
# Performance

## Performance of Swin-DINO on validation set

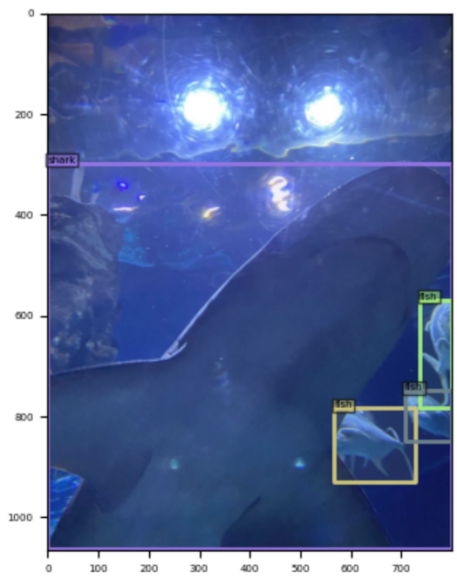
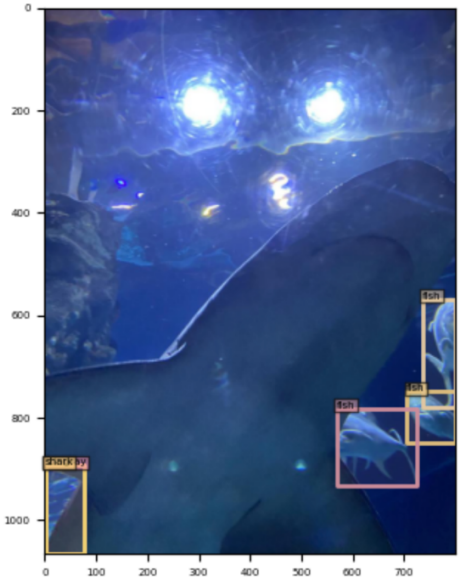
| AP | $AP_{50}$ | $AP_{75}$ |
|---|---|---|
| 0.586 | 0.859 | 0.614 |

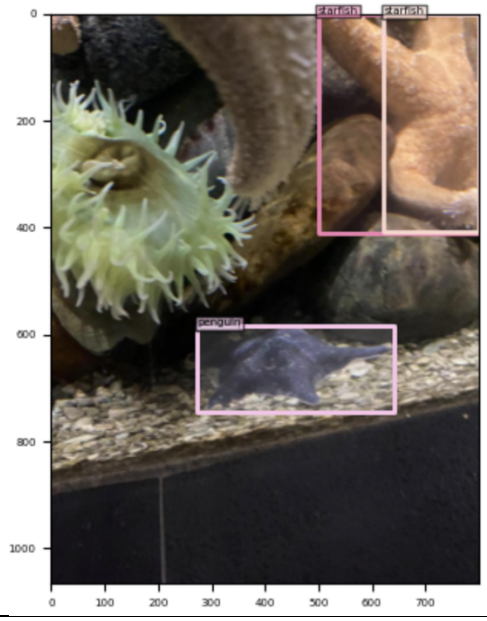
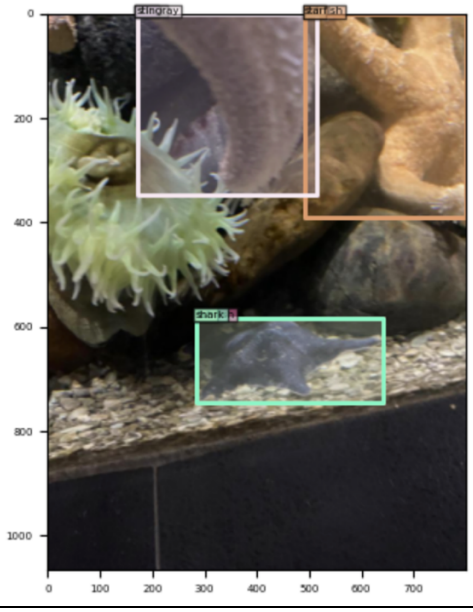
```
IoU metric: bbox
 Average Precision  (AP) @[ IoU=0.50:0.95 | area=   all | maxDets=100 ] = 0.586
 Average Precision  (AP) @[ IoU=0.50      | area=   all | maxDets=100 ] = 0.859
 Average Precision  (AP) @[ IoU=0.75      | area=   all | maxDets=100 ] = 0.614
 Average Precision  (AP) @[ IoU=0.50:0.95 | area= small | maxDets=100 ] = 0.226
 Average Precision  (AP) @[ IoU=0.50:0.95 | area=medium | maxDets=100 ] = 0.466
 Average Precision  (AP) @[ IoU=0.50:0.95 | area= large | maxDets=100 ] = 0.723
 Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets=  1 ] = 0.269
 Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets= 10 ] = 0.595
 Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets=100 ] = 0.715
 Average Recall     (AR) @[ IoU=0.50:0.95 | area= small | maxDets=100 ] = 0.468
 Average Recall     (AR) @[ IoU=0.50:0.95 | area=medium | maxDets=100 ] = 0.637
 Average Recall     (AR) @[ IoU=0.50:0.95 | area= large | maxDets=100 ] = 0.812
Training time 1:43:00
```

## Performance on validation set

| Swin + DINO | ResNet50 + DINO |
|---|---|
|  |  |

- Visualization
  - On Testing Set
    - Link : https://github.com/irisowo/CVPDL-HW1/tree/main/DINO/figs/imgs
    - Preview

| | Success Case (Swin) | Failure Case (Res24+24) |
|---|---|---|
| 5<br>shark |  |  |

| | Failure Case (Swin) | Failure Case (Res24+24) |
|---|---|---|
| 5<br>shark |  |  |