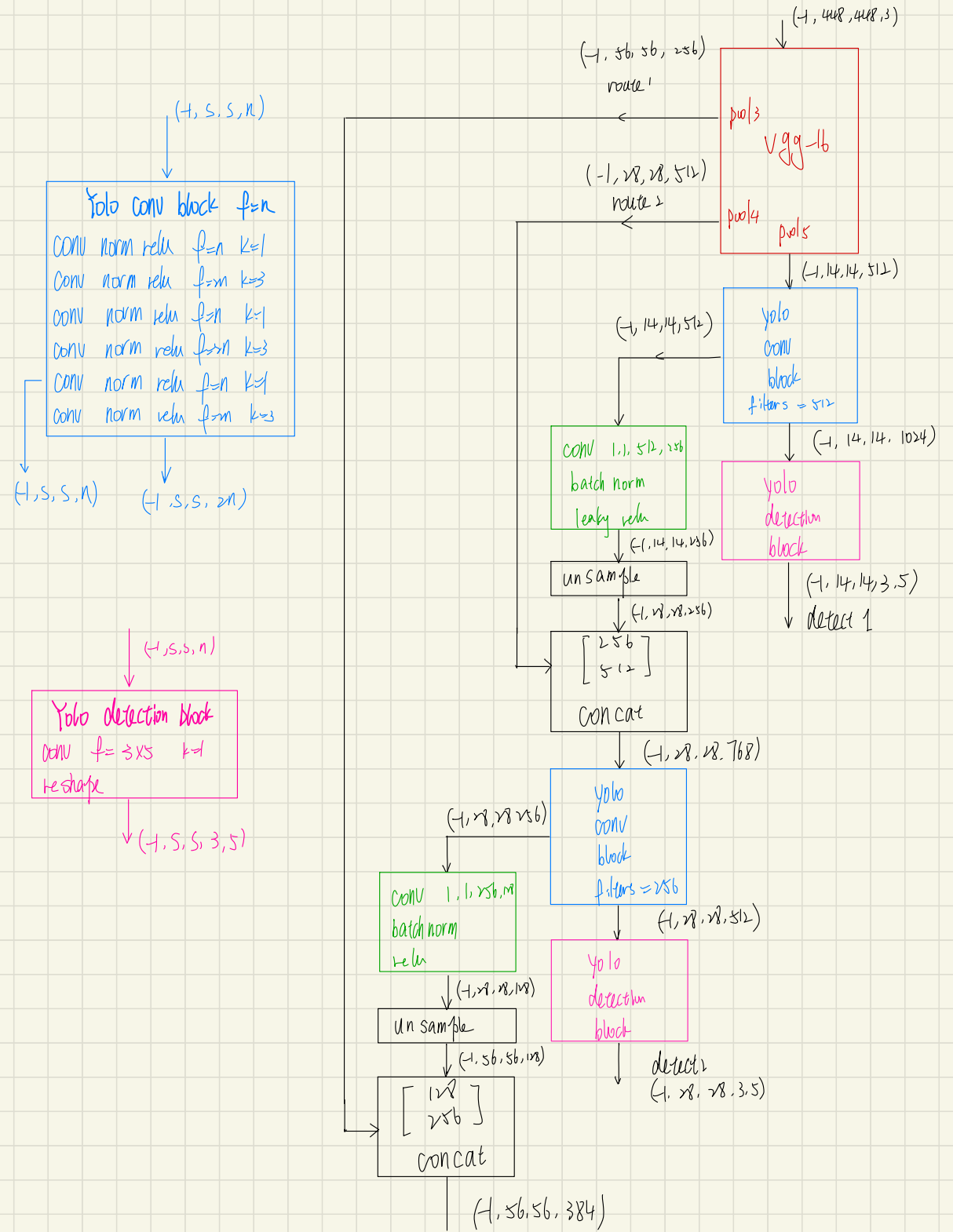
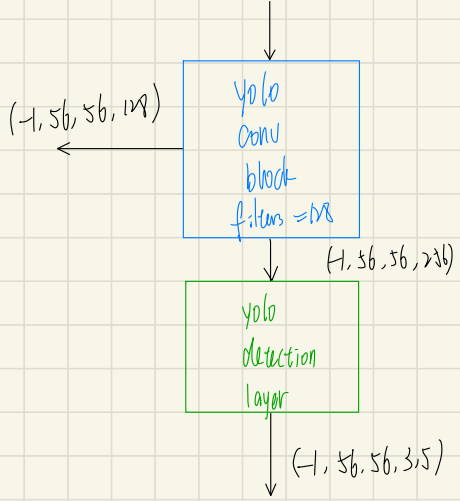
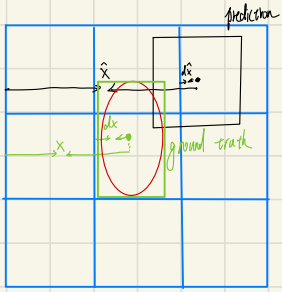

(-1, 416, 416, 3)
(-1, 52, 52, 256)
route 1
pool3
Vgg-16
(-1, 26, 26, 512)
route 2
pool4
pool5
(-1, 13, 13, 512)
yolo conv block
filters = 512
(-1, 13, 13, 1024)
yolo detection block
(-1, 13, 13, 3, 5)
detect 1
(-1, 13, 13, 512)
conv 1, 1, 512, /2
batch norm
leaky relu
(-1, 13, 13, 256)
unsample
(-1, 26, 26, 256)
[256, 512]
concat
(-1, 26, 26, 768)
yolo conv block
filters = 256
(-1, 26, 26, 512)
yolo detection block
detect 2
(-1, 26, 26, 3, 5)
(-1, 26, 26, 256)
conv 1, 1, 256, /1
batch norm
relu
(-1, 26, 26, 128)
unsample
(-1, 52, 52, 128)
[128, 256]
concat
(-1, 52, 52, 384)
(-1, S, S, n)
Yolo conv block f=n
conv norm relu f=n k=1
conv norm relu f=n k=3
conv norm relu f=n k=1
conv norm relu f=n k=3
conv norm relu f=n k=1
conv norm relu f=n k=3
(-1, S, S, n)
(-1, S, S, 2n)
(-1, S, S, n)
Yolo detection block
conv f=3x5 k=1
reshape
(-1, S, S, 3, 5)

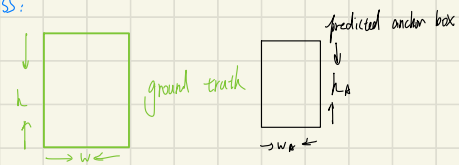


coordinate loss:



$$L_{coord} = (x-\hat{x})^2 + (y-\hat{y})^2$$

size loss:

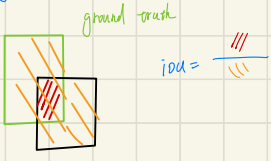


$$L_{size} = \sum_{k \in A} \mathbb{1}\{\text{face in } k\} \cdot \left\{ (h - p_h \cdot h_a)^2 + (w - p_w \cdot w_a)^2 \right\}$$

A are all anchor boxes

obj loss: (when there is a face in the anchor box in the grid cell)

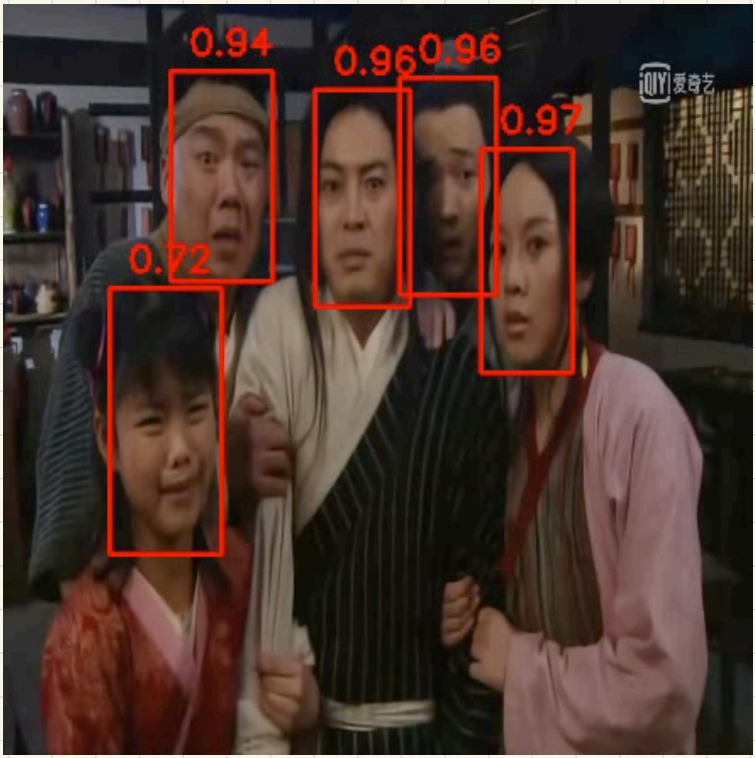
$$L_{obj} = \sum_i \sum_j \sum_{k \in A} \mathbb{1}\{\text{face in } k\} (\hat{p}_{obj} - iou)^2$$



noobj loss: (false positive)

$$L_{noobj} = \sum_i \sum_j \sum_{k \in A} \mathbb{1}\{\text{no face in } k\} (\hat{p}_{obj})^2$$

$$L = \lambda_{coord}(L_{coord} + L_{size}) + \lambda_{obj} \cdot L_{obj} + \lambda_{noobj} \cdot L_{noobj}$$


0.94
0.96
0.96
0.97
0.72
iQIYI 爱奇艺