GigaVision Challenge

When Gigapixel Videography Meets Computer Vision

Track： Gigadetection

Team name： HIEVE

Catalogue

1. Team Introduction
2. Task Analysis
   - a. Scene background complexity
   - b. Target distribution variance
3. Solution and Innovation
   - a. Data Preprocess
   - b. Model Details
   - c. Data Postprocess
4. Results & Visualization
5. Conclusion

GIGAVISION

Catalogue

1. Team Introduction

2. Task Analysis

a. Scene background complexity

b. Target distribution variance

3. Solution and Innovation

a. Data Preprocess

b. Model Details

c. Data Postprocess

4. Results & Visualization

5. Conclusion

Team Introduction

HIEVE(human in event)

团队成员：彭嘉淇

上海交通大学电子信息与电气工程学院
电子工程系本科生，大四

指导老师：林巍峣

上海交通大学电子信息与电气工程学院
电子工程系教授

GIGAVISION

Catalogue

1. Team Introduction
2. Task Analysis
    a. Scene background complexity
    b. Target distribution variance
3. Solution and Innovation
    a. Data Preprocess
    b. Model Details
    c. Data Postprocess
4. Results & Visualization
5. Conclusion

Task Analysis

PANDA：

(1) globally wide field-of-view where visible area may beyond 1 km^2,
(2) locally high resolution details with gigapixel-level spatial resolution,
(3) temporally long-term crowd activities with 43.7k frames in total,
(4) real-world scenes with abundant diversities in human attributes, behavioral patterns, scale, density, occlusion, and interaction.

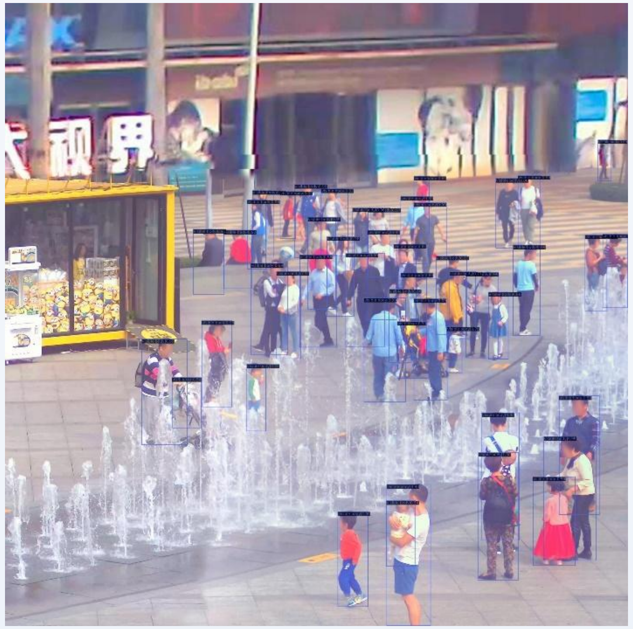
(a)

(b)

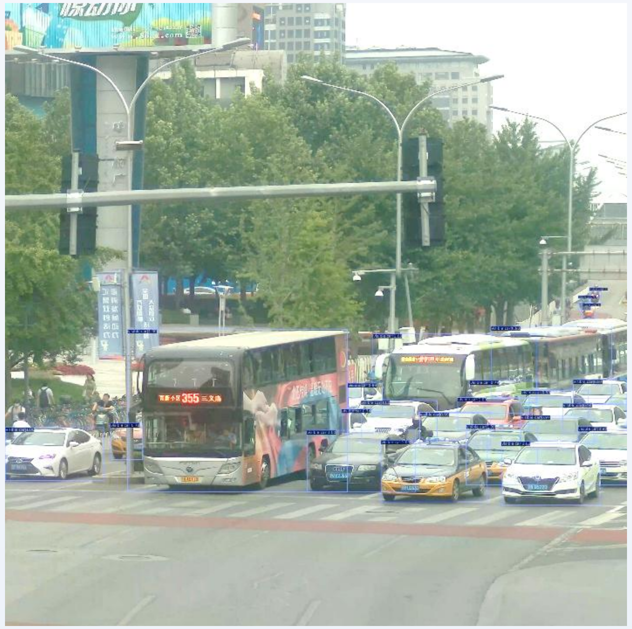
Task Analysis

Object Detection for Gigapixel Images:

The track aims to detect pedestrians and vehicles in gigapixel images. Most of the existing detection algorithms are designed based on data sets with low resolution such as MSCOCO (about 600×400 pixels), and the performance is reduced due to significant differences in scale, pose and occlusion.



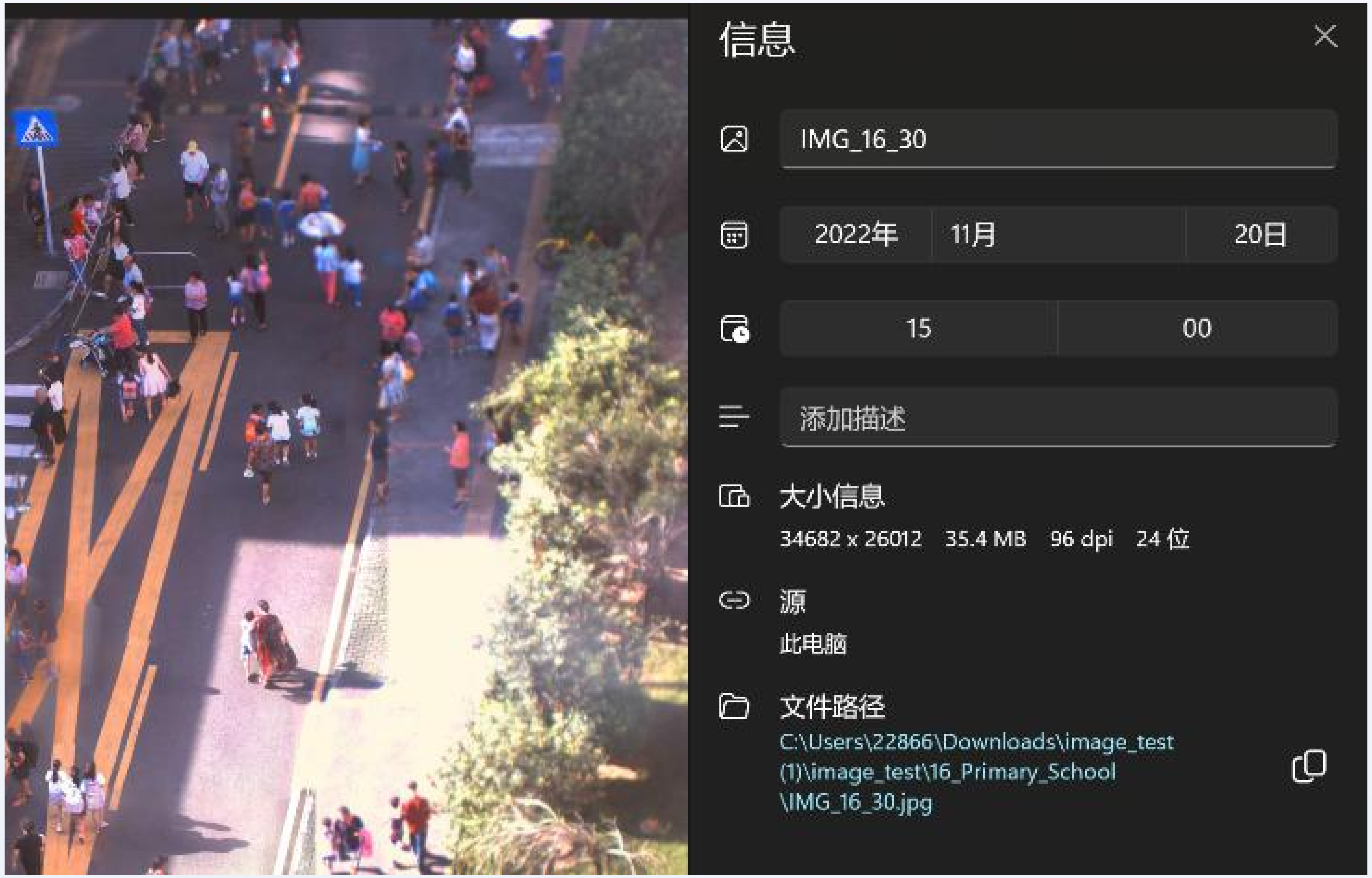
pedestrians



vehicles

Task Analysis

Challenge 1: gigapixel resolution, near billion pixels (essential feature)



信息

IMG_16_30

2022年 11月 20日

15 00

添加描述

大小信息

34682 x 26012 35.4 MB 96 dpi 24 位

源

此电脑

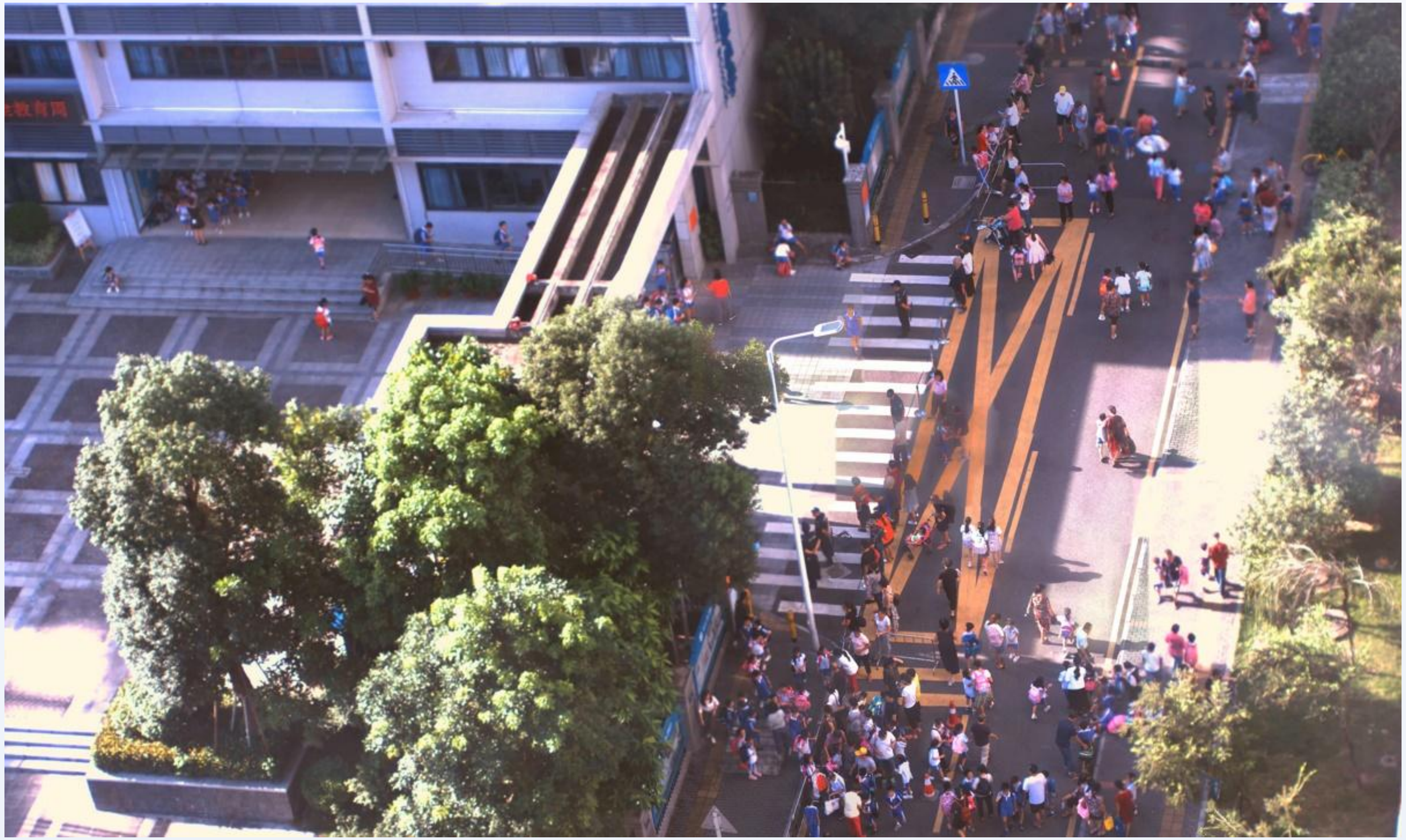
文件路径

C:\Users\22866\Downloads\image_test (1)\image_test\16_Primary_School \IMG_16_30.jpg

Task Analysis

Challenge 2： background complexity

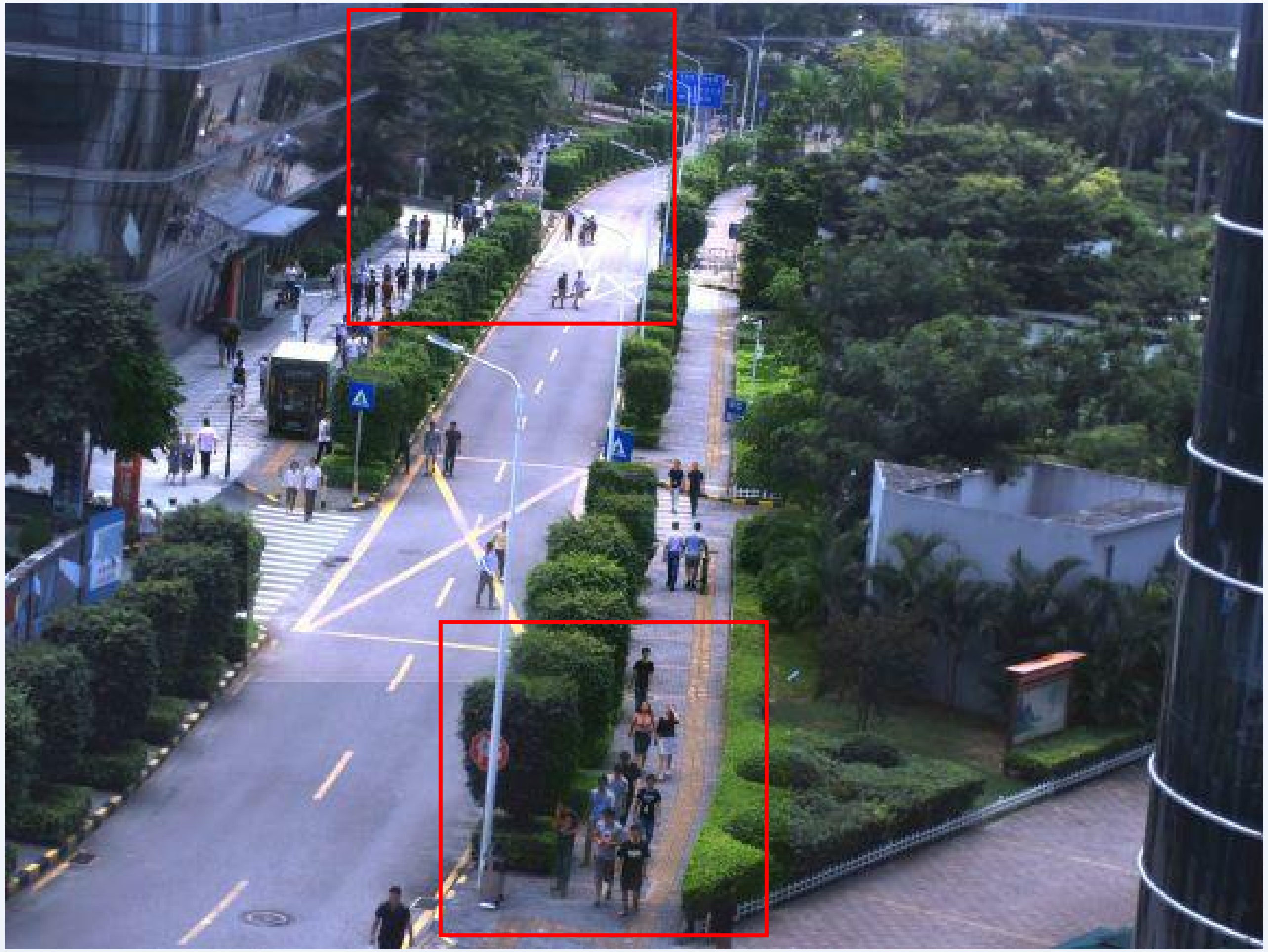
- Immobility
- Diversity

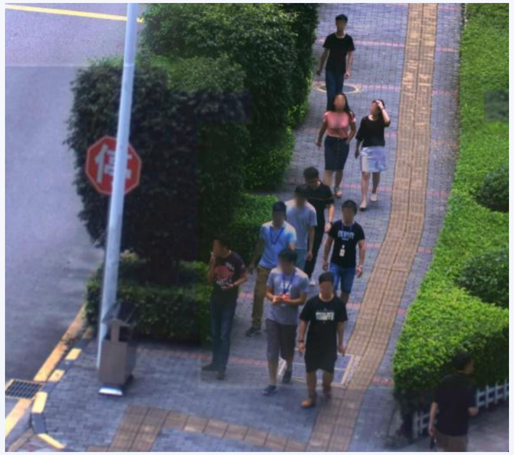


Task Analysis

Challenge 3： distribution variance

- location distribution
- scale distribution





GIGAVISION

Catalogue

1. Team Introduction
2. Task Analysis
   a. Scene background complexity
   b. Target distribution variance
3. Solution and Innovation
   a. Data Preprocess
   b. Model Details
   c. Data Postprocess
4. Results & Visualization
5. Conclusion

Data Preprocess

Data Preprocess

Window partitioning

In the training phase, the training set images are divided into smaller windows. In order to make the model pay more attention to the target, the background images without the target are removed.

1500x1500(part)

3000x3000

6000x6000

12000x12000

Data Preprocess

Background generalization

Since the training dataset only has 13 scenes, which is easy to overfit. Other datasets are added to provide more abundant background information to improving the generalization ability of the model.

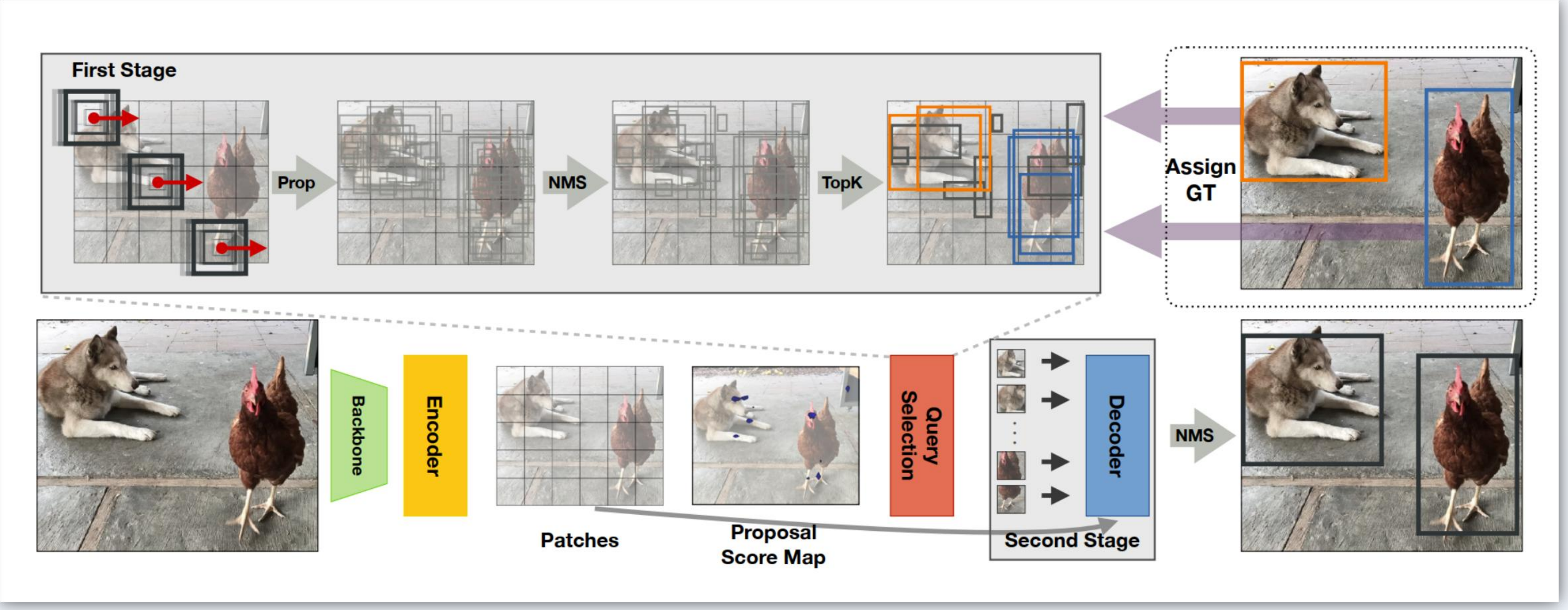
dataset	images	annotatioms
MOT17	5316	112297
CrowdHuman	15000	438792
CUHK-SYSU	6062	32203
PRW	5908	21462

Model Details

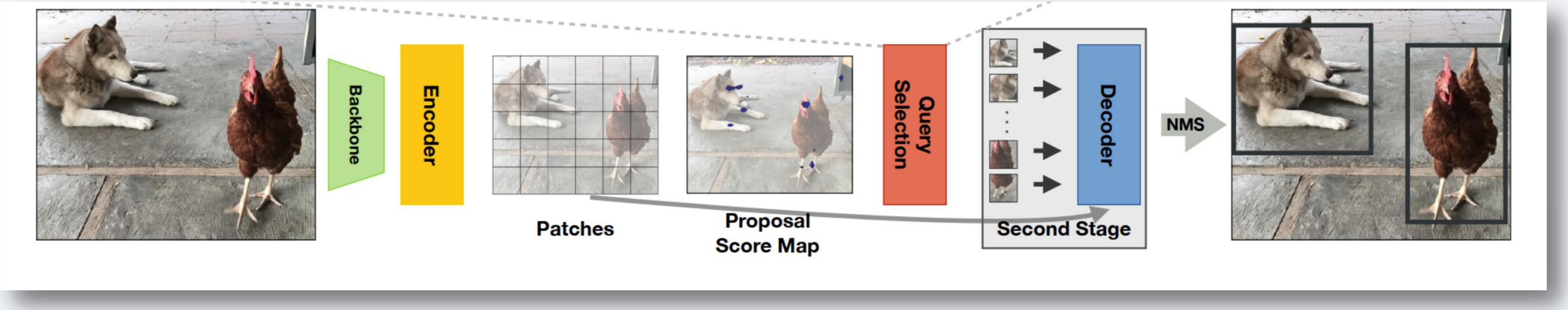
Model Details

DETA

A two-stage end-to-end detector with the following network structure:



Model Details

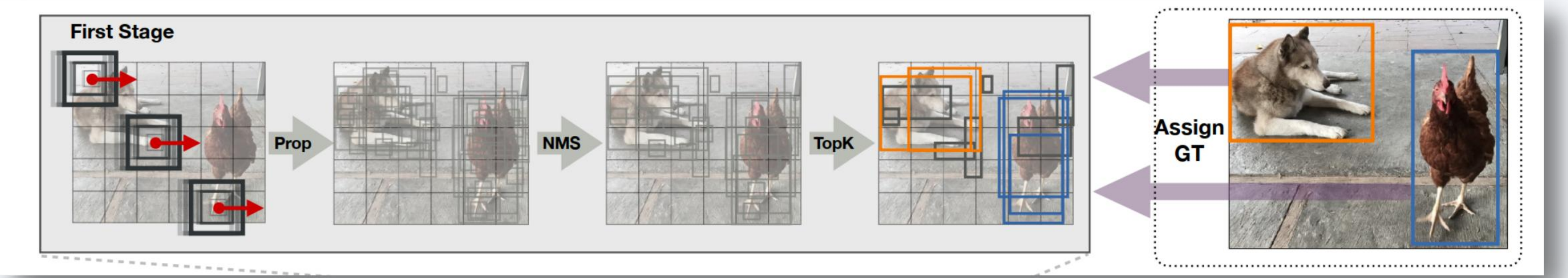


Two stage assignments :

In the first stage, the pre-set boxes is as a queries to the transformer encoder to extract image features and generate a score map, and then the classification category and prediction boxes are predicted through the decoder to match with the ground truth.

In the second stage, the image features extracted from the encoder in the first stage are used as queries to further refine the result. Using the dynamic query can further improve the accuracy of the model and accelerate the convergence of the model.

Model Details



$$\sigma_i^{prop} = \begin{cases} \hat{k} = argmax_k IOU\left(b_i^{prop}, b_k\right),\ if\ IOU\left(b_i^{prop}, b_k\right) \geq \tau_k^n \\ \qquad\qquad\qquad\qquad or C_{\max}^{prop}\left(i, \hat{k}\right) \\ \emptyset, \qquad\qquad otherwise \end{cases}$$

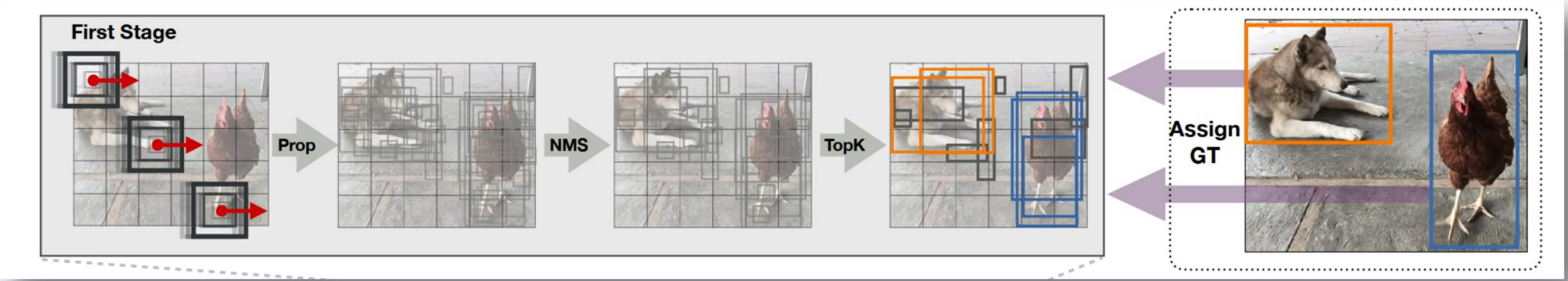
Condition:

$$IOU\left(b_i^{prop}, b_k\right) \geq \tau_k^n$$

$$C_{\max}^{prop}\left(i, \hat{k}\right) = \bigcap_{j \neq i} \{IOU(b_i^{init}, b_k) \geq IOU(b_j^{init}, b_k)\}$$

One-to-many matching, that is, one ground truth matches multiple prediction boxes, which brings more positive samples, accelerates the training process.

Model Details



$$\sigma_i^{prop} = \begin{cases} \hat{k} = argmax_k IOU\left(b_i^{prop}, b_k\right),\ if\ IOU\left(b_i^{prop}, b_k\right) \geq \tau_k^n \\ \qquad\qquad\qquad\qquad or C_{\max}^{prop}\ (i, \hat{k}) \\ \emptyset, \qquad\qquad otherwise \end{cases}$$

dynamic threshold:

$$\tau_k^n = \max\left(\tau,\ \mu_k^n\right)$$

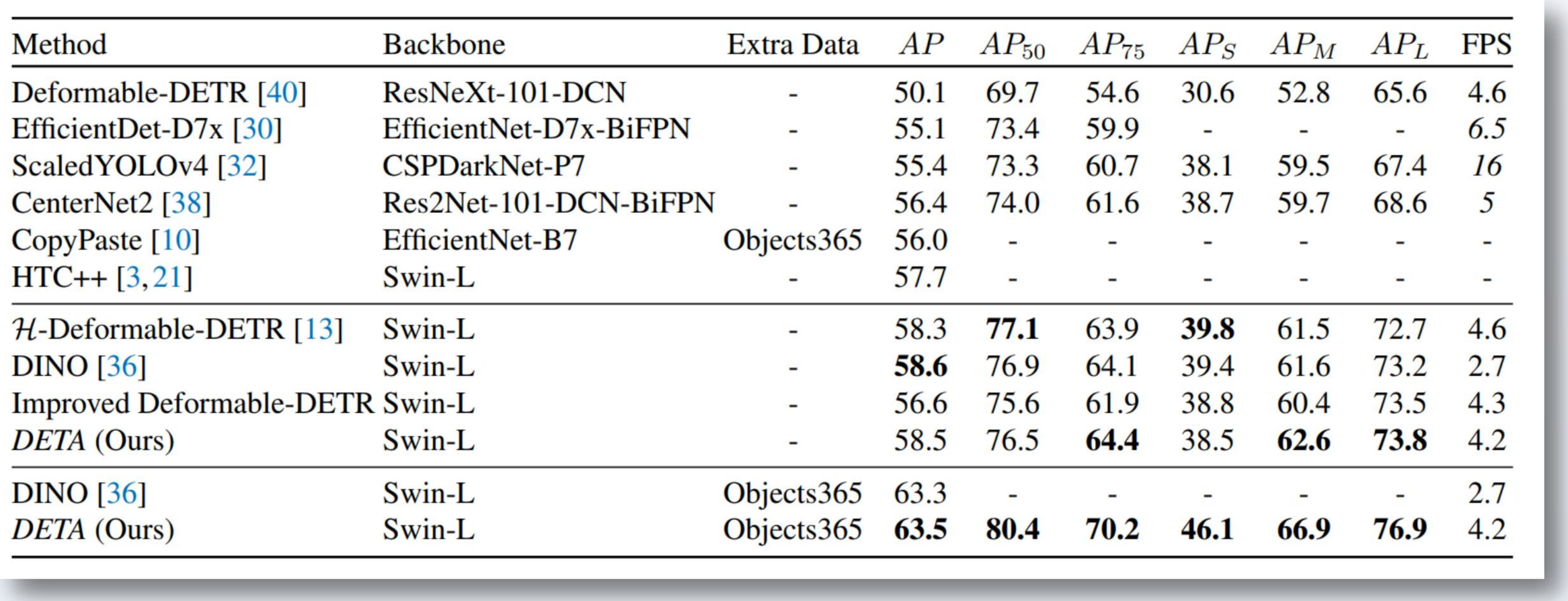
Where μ_k^n : the n-th highest IoU for annotation k

A **dynamic balance** strategy is used to adaptively adjust the iou threshold in the matching process to prevent the performance from being affected by the large gap between the matching number of large targets and small targets, improving the performance for small targets.

Model Details

DETA

The Swin-L is used as the backbone, and the weights of DETA(Swin-L) trained on COCO for 24 epochs are used as the pre-trained model.



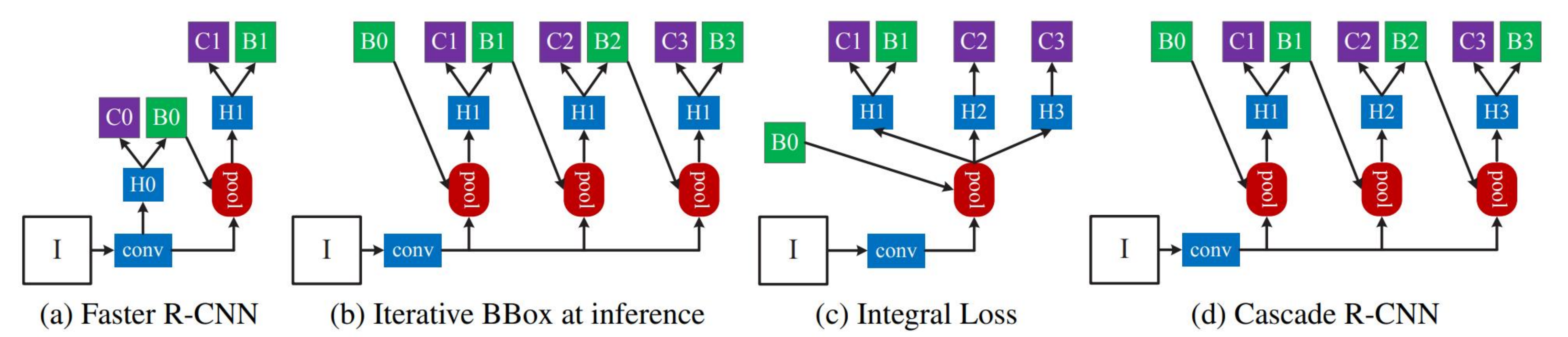
Method	Backbone	Extra Data	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS
Deformable-DETR [40]	ResNeXt-101-DCN	-	50.1	69.7	54.6	30.6	52.8	65.6	4.6
EfficientDet-D7x [30]	EfficientNet-D7x-BiFPN	-	55.1	73.4	59.9	-	-	-	*6.5*
ScaledYOLOv4 [32]	CSPDarkNet-P7	-	55.4	73.3	60.7	38.1	59.5	67.4	*16*
CenterNet2 [38]	Res2Net-101-DCN-BiFPN	-	56.4	74.0	61.6	38.7	59.7	68.6	*5*
CopyPaste [10]	EfficientNet-B7	Objects365	56.0	-	-	-	-	-	-
HTC++ [3,21]	Swin-L	-	57.7	-	-	-	-	-	-
$\mathcal{H}$-Deformable-DETR [13]	Swin-L	-	58.3	**77.1**	63.9	**39.8**	61.5	72.7	4.6
DINO [36]	Swin-L	-	**58.6**	76.9	64.1	39.4	61.6	73.2	2.7
Improved Deformable-DETR	Swin-L	-	56.6	75.6	61.9	38.8	60.4	73.5	4.3
DETA (Ours)	Swin-L	-	58.5	76.5	**64.4**	38.5	**62.6**	**73.8**	4.2
DINO [36]	Swin-L	Objects365	63.3	-	-	-	-	-	2.7
DETA (Ours)	Swin-L	Objects365	**63.5**	**80.4**	**70.2**	**46.1**	**66.9**	**76.9**	4.2

Model Details

other models

Cascade RCNN

Cascade R-CNN consists of a sequence of detectors trained with increasing IoU thresholds, to be sequentially more selective against close false positives.



(a) Faster R-CNN (b) Iterative BBox at inference (c) Integral Loss (d) Cascade R-CNN

yolov7

1. Extended efficient layer aggregation networks: uses expand, shuffle, merge cardinality to achieve the ability to continuously enhance the learning ability of the network without destroying the original gradient path.

2. Extra auxiliary head: use a coarse-to-fine lead head guided label assigner, coarse label is generated by allowing more grids to be treated as positive target by relaxing the constraints of the positive sample assignment process.

3. Planned re-parameterized convolution: merge multiple computational modules into one at inference stage.

4. Model scaling: adjust some attributes of the model and generate models of different scales to meet the needs of different inference speeds.

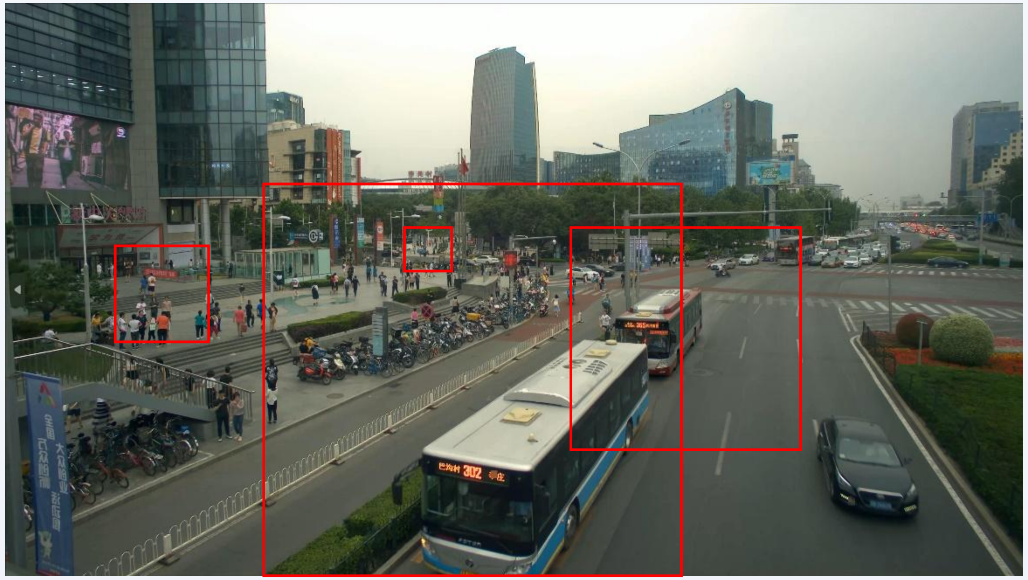
Data Postprocess

Data Postprocess

Multi-scale testing

small scale: can't cover large targets;

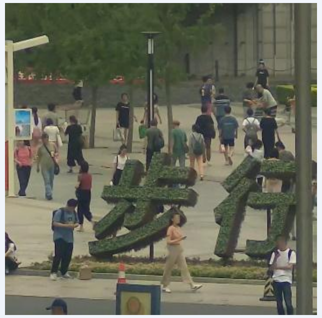
large scale: easy to ignore small targets;



12000x12000

6000x6000

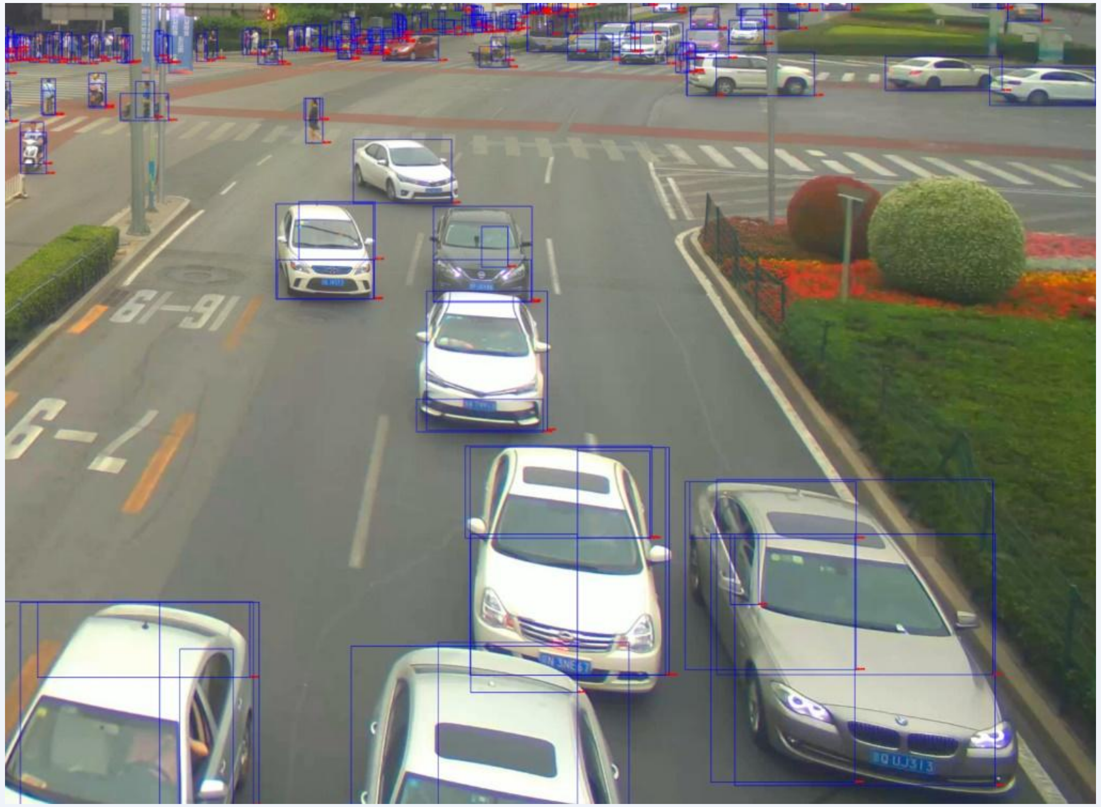
3000x3000



1500x1500

Data Postprocess

Boundary judgment

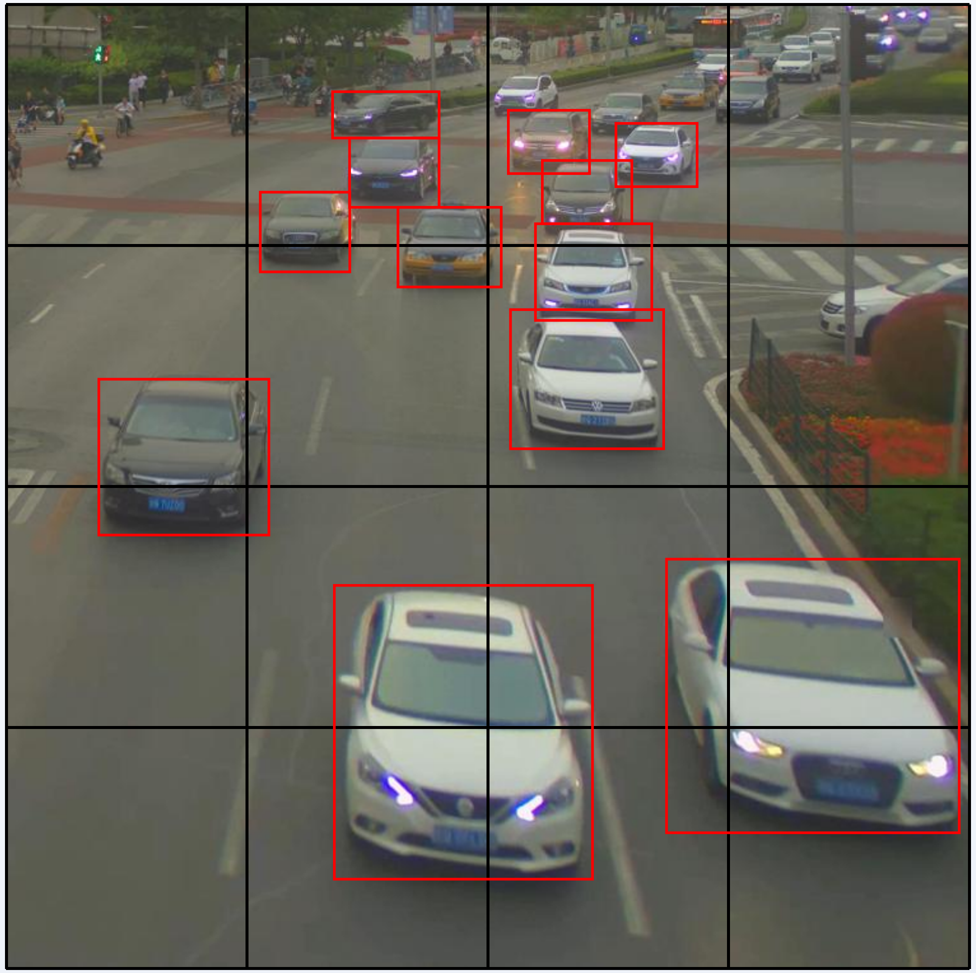


Solution 1:
directly remove the predicted box near the boundary.

Data Postprocess

Boundary judgment

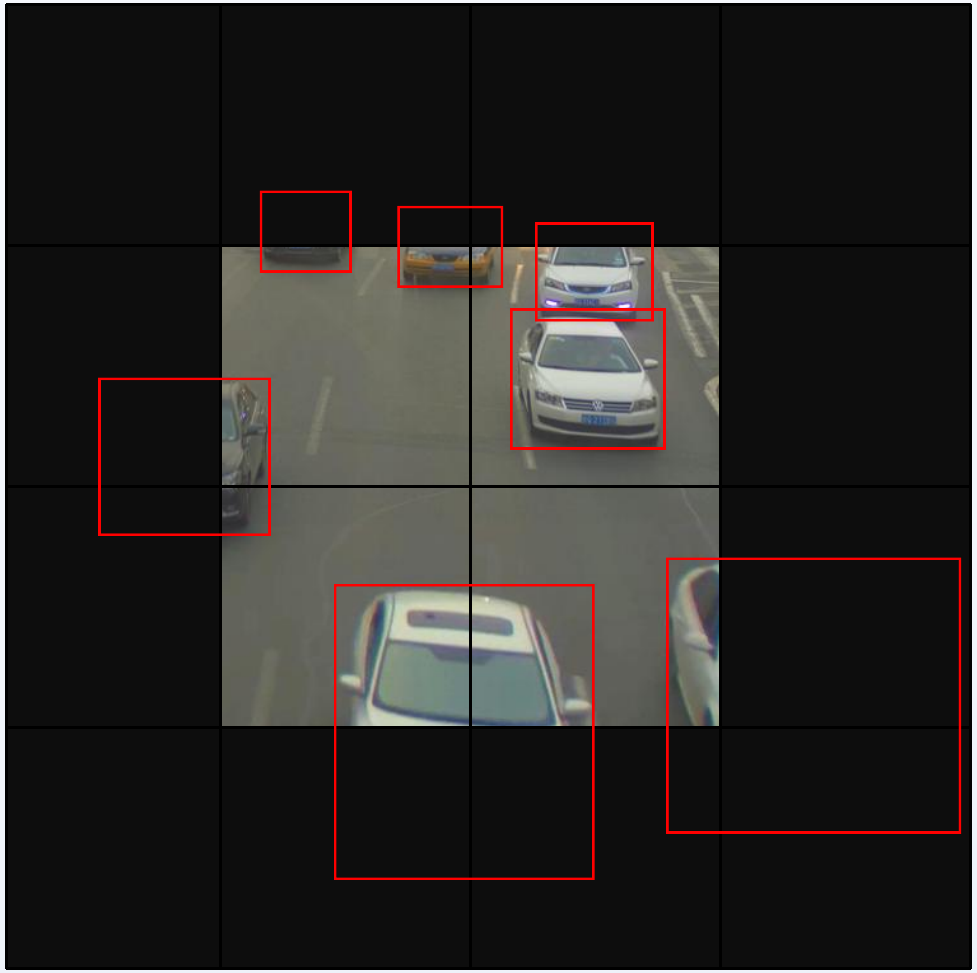
Solution 2



(1) Full image detection

Data Postprocess

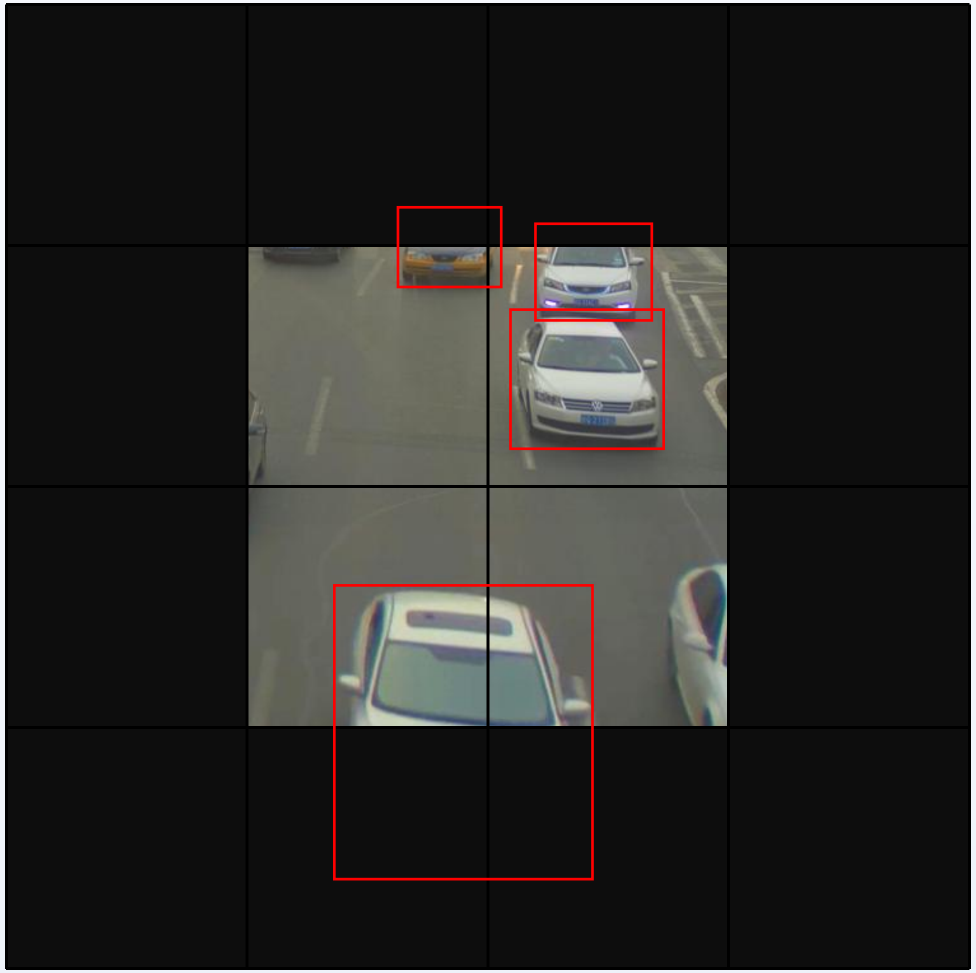
Boundary judgment



(2) keep the middle boxes

Data Postprocess

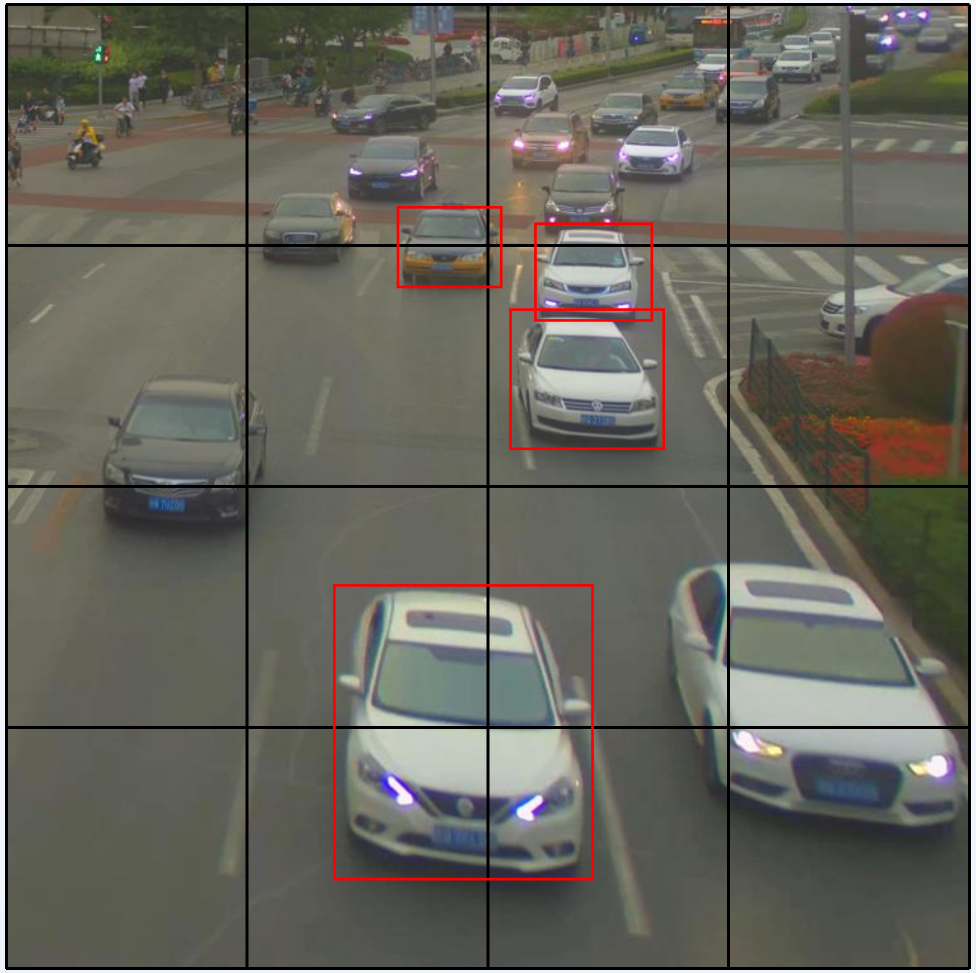
Boundary judgment



(3) remove boxes with IOF less than 0.5

Data Postprocess

Boundary judgment



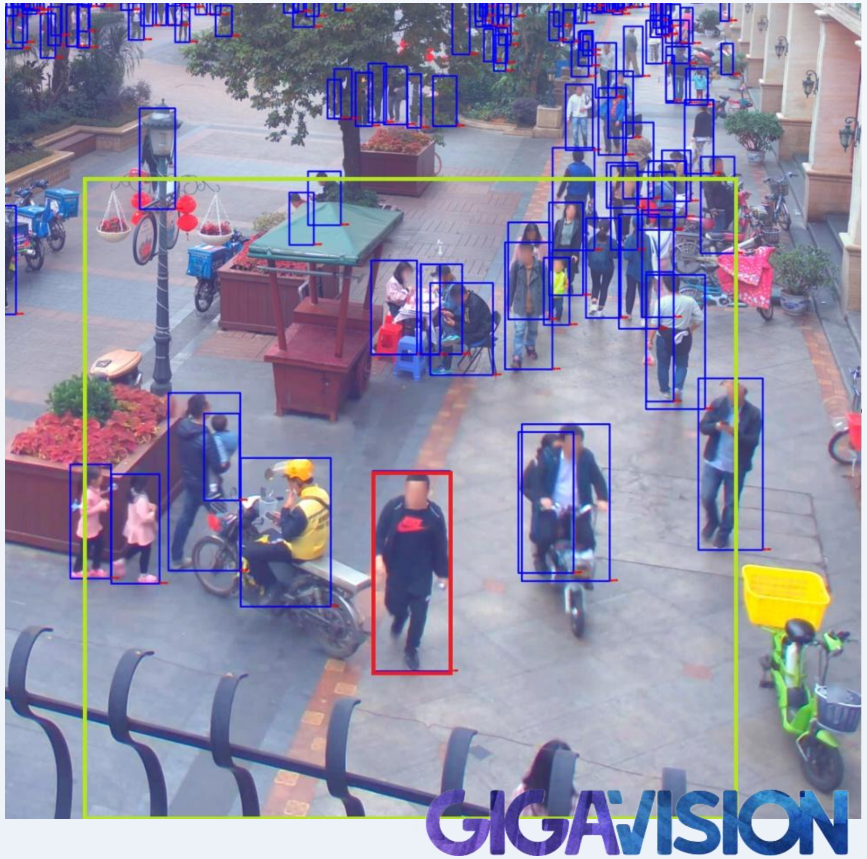
(4) final result

Data Postprocess

Iterative inference:

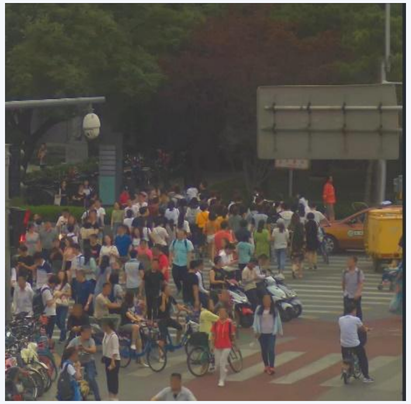
According to the detection results of the first stage, box size can be represented as the scale of the scene. Keep the size of the sub image in an appropriate proportion with the size of all objects in the sub image, and recut the test set images.

```
for img in all_imgs:
    for box in img:
        if marked:
            continue
        cx, cy, w, h <= box
        scale = max(h, w)
        side = scale*stride
        crop_region = (cx, cy, side, side)
        for i in N:
                adjacent_target = iou(crop_region, boxes) > thr
                scale = mean([max(h, w) for box in adjacent_target])
                side = scale*stride
                crop_region = (cx, cy, side, side)
        mark overlap_box
        crop image
```



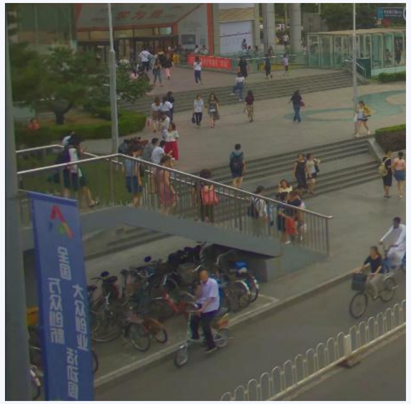
Data Postprocess

1370x1370

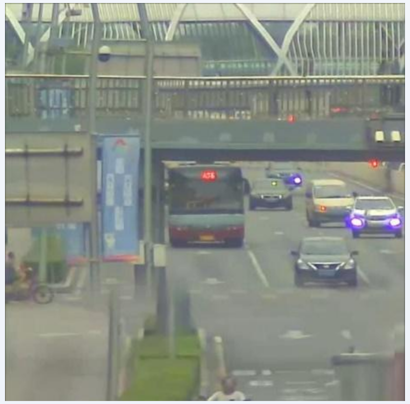


2130x2130

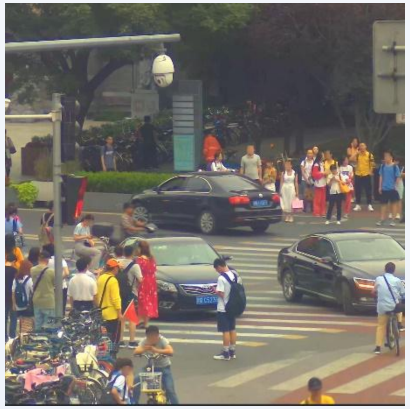
3745x3745



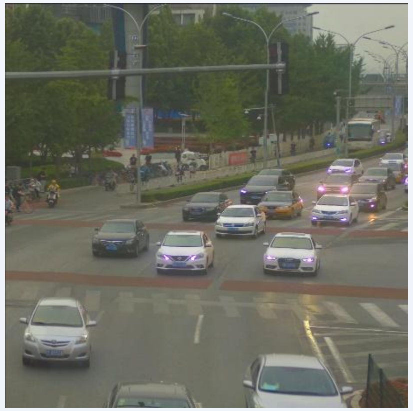
8000x8000



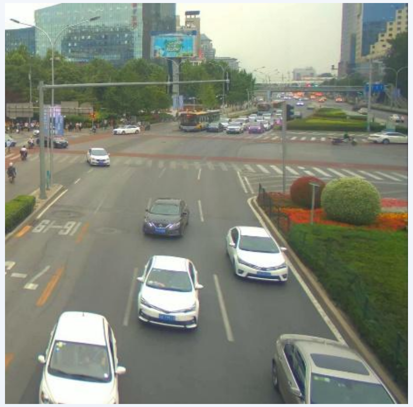
706x706



1710x1710



4066x4066



11383x11383

GIGAVISION

Data Postprocess

Iterative inference:

First stage（yolov7）:

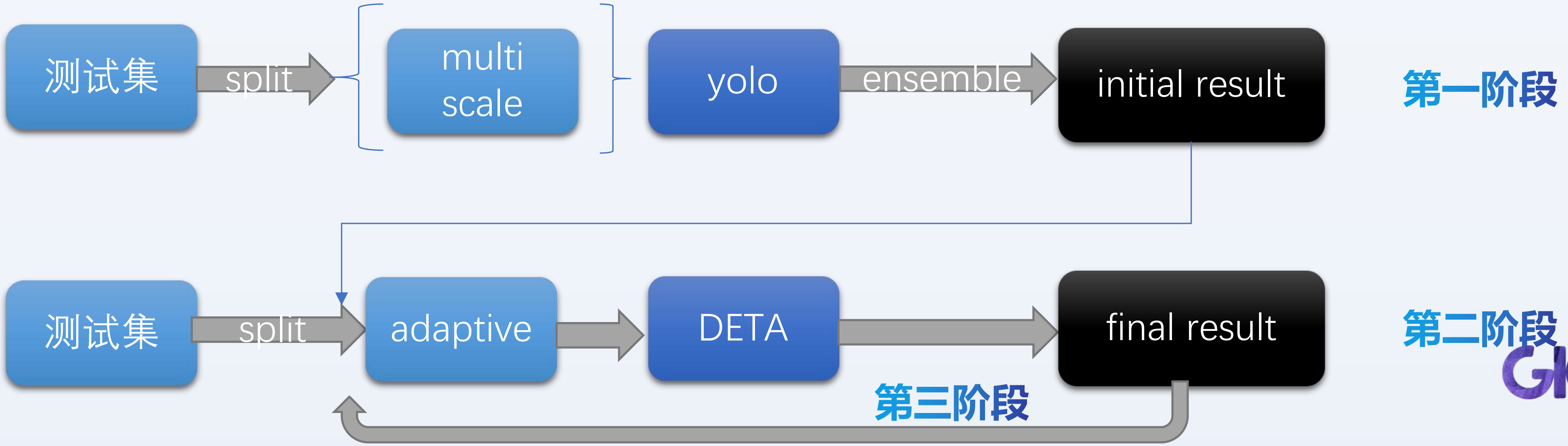
yolov7 is efficient due to using techniques such as convolution reparametrization to achieve high performance in speed and accuracy. The goal of the first stage here is not to detect objects accurately, but to cover all objects in the image and estimate their size correctly.

Second stage （DETA)

After Iterative inferencing, only the region with the target is retained, and the size of the sub image is optimized so that the target scale distribution is more consistent, which alleviates the scale variance, reduces the number of images to inference, and improves the speed and accuracy of detection.

Third stage（DETA)

Iterate again according to its result until convergence to a stable value.



Catalogue

1. Team Introduction
2. Task Analysis
    a. Scene background complexity
    b. Target distribution variance
3. Solution and Innovation
    a. Data Preprocess
    b. Model Details
    c. Data Postprocess
4. Results & Visualization
5. Conclusion

Results & Visualization

compare the results of the three different methods tried on the PANDA dataset

Method	AP	AR	score
Cascade RCNN	0.53	0.6120	0.569
Yolov7	0.67	0.7198	0.693
DETA	0.74	0.8116	0.774

compare the results of iterative inference and single-scale inference

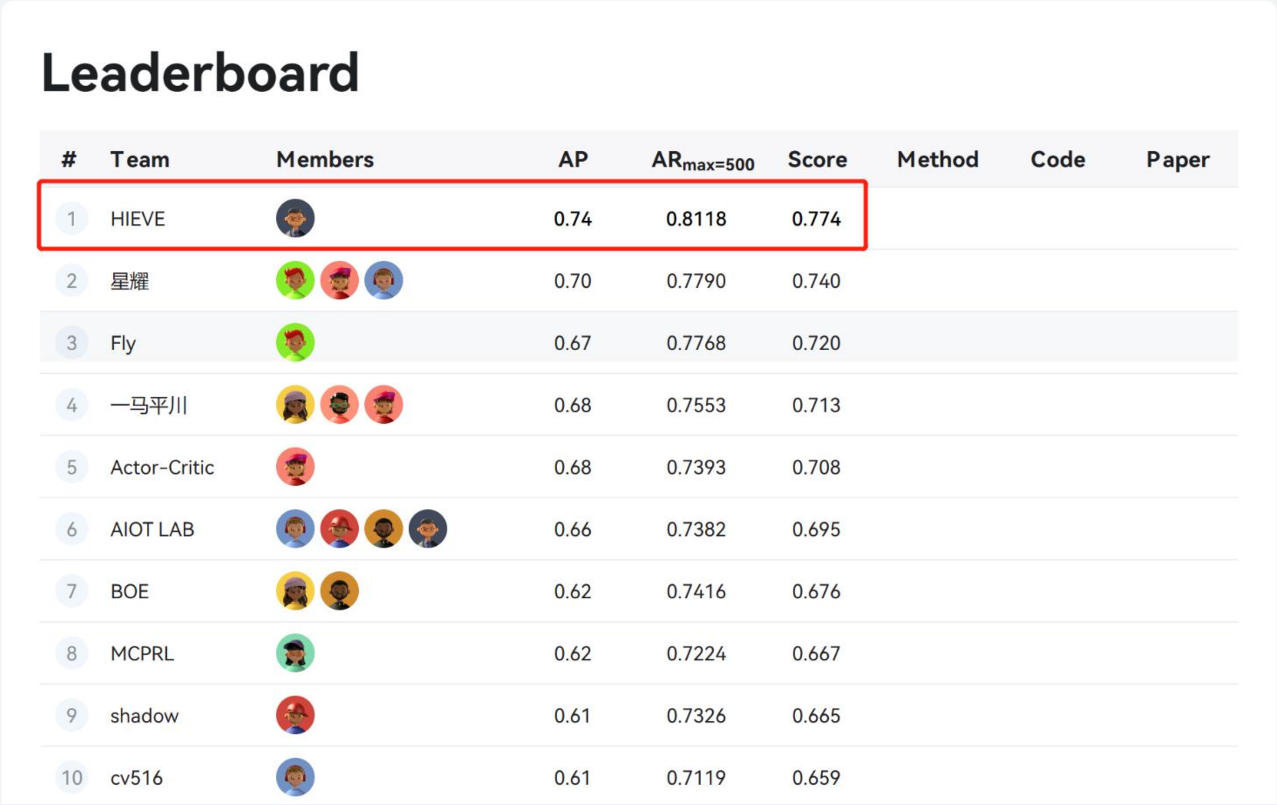
Method	AP	AR	score
single-scale	0.72	0.7848	0.750
迭代推理一次	0.73	0.8010	0.763
迭代推理至收敛	0.74	0.8118	0.774

GIGAVISION

Results & Visualization

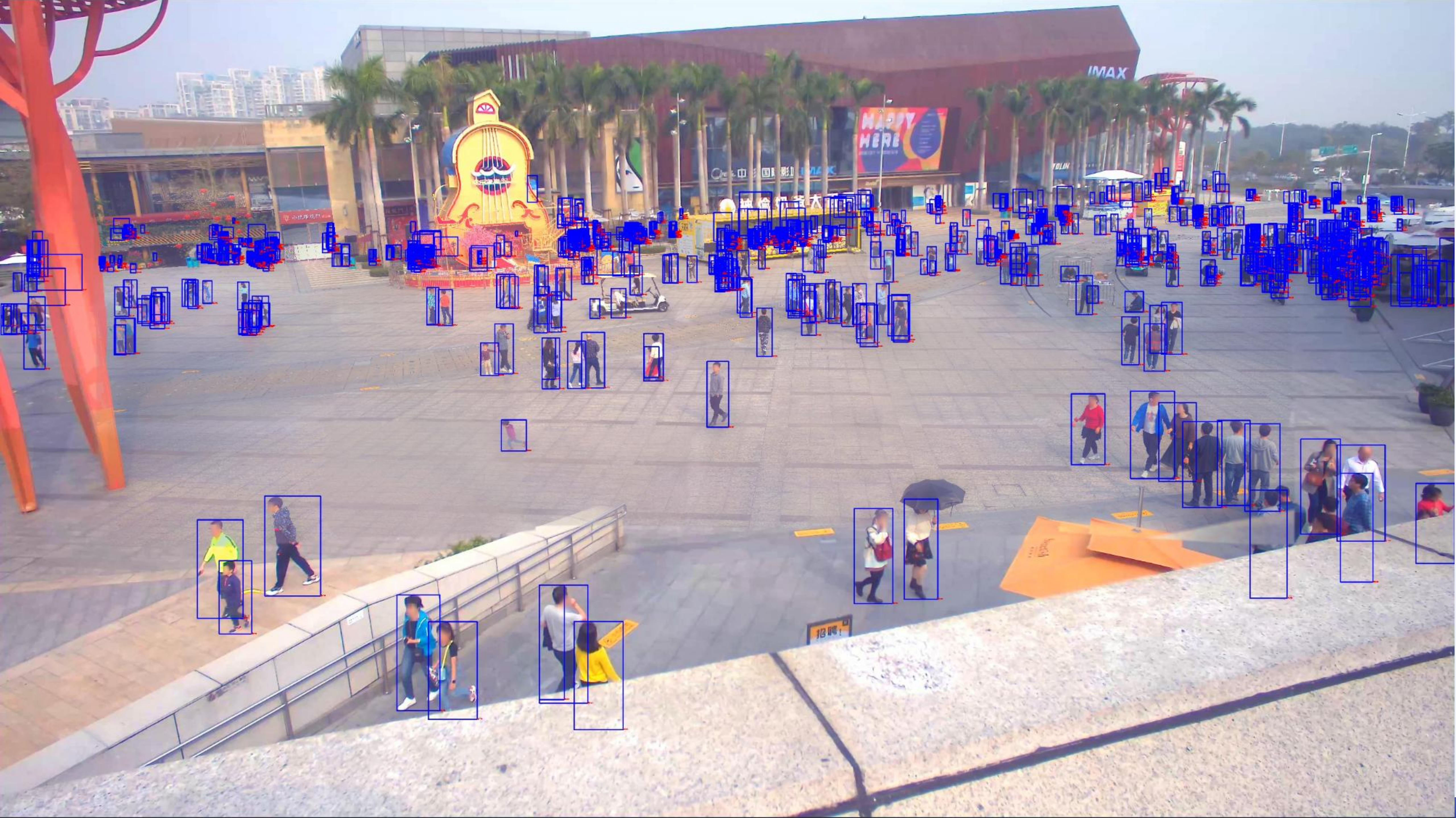
Final result

Method	AP	AR	score
ours	0.74	0.8118	0.774

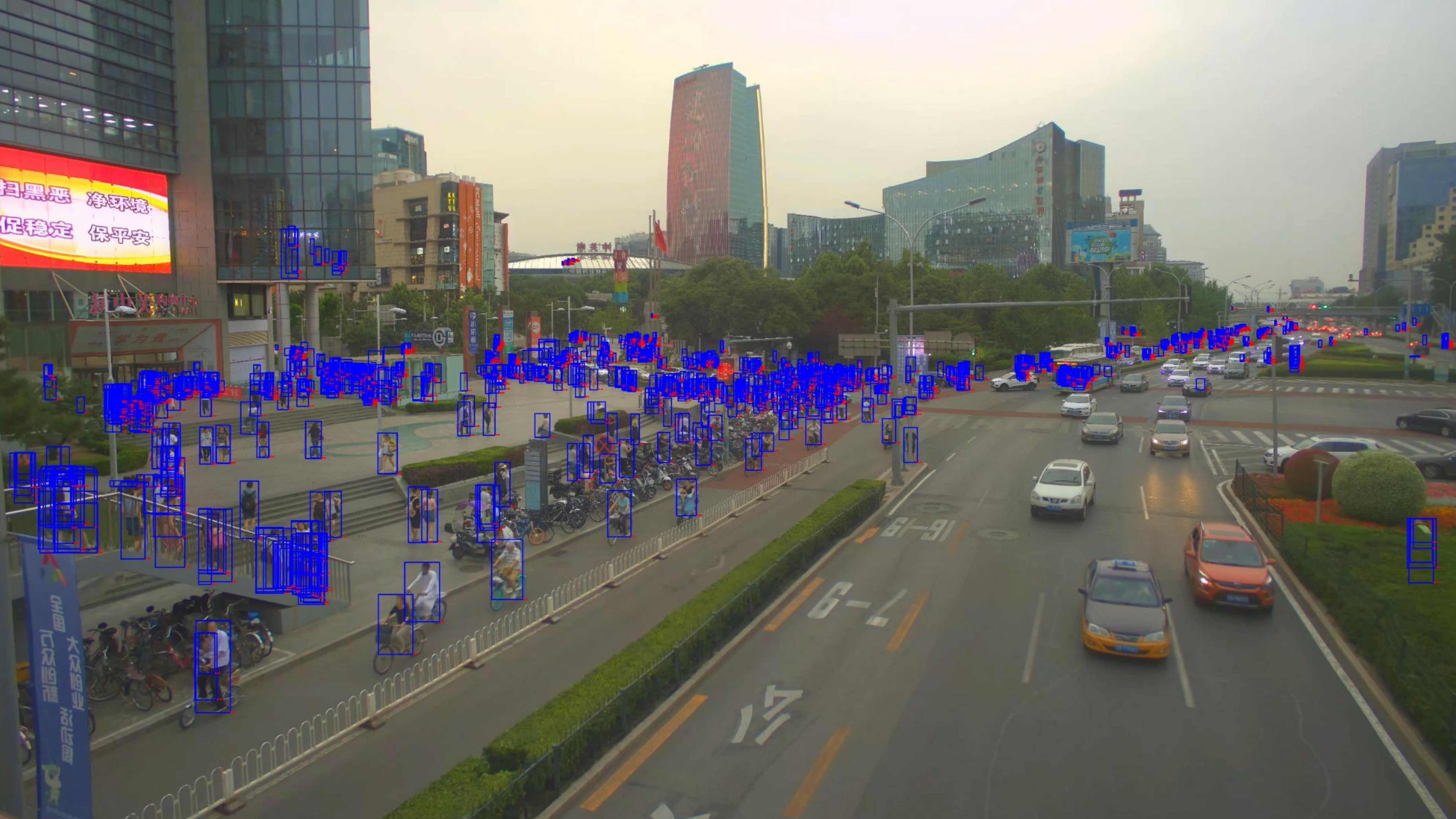


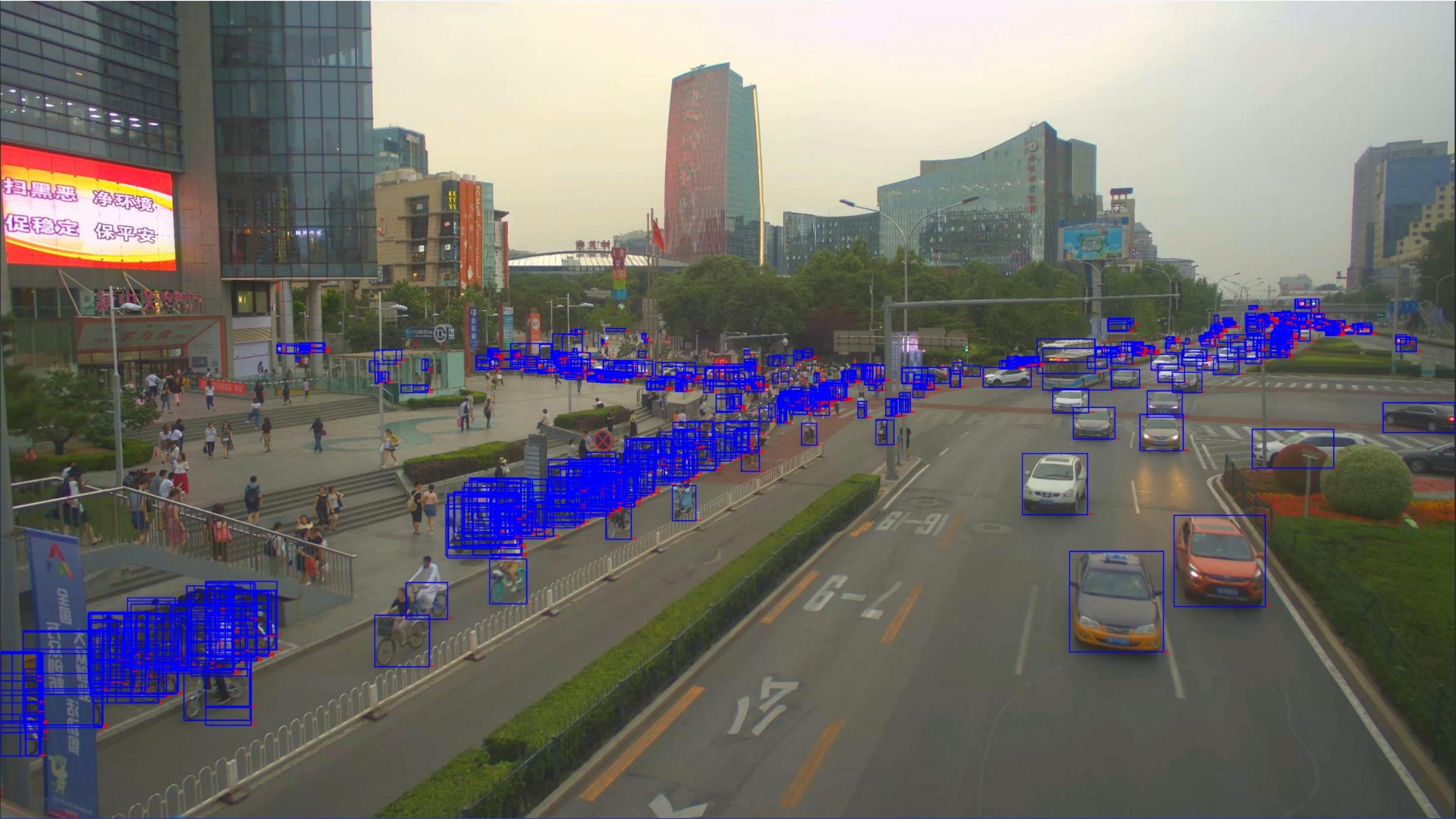
Leaderboard

#	Team	Members	AP	$AR_{max=500}$	Score	Method	Code	Paper
1	HIEVE		0.74	0.8118	0.774			
2	星耀		0.70	0.7790	0.740			
3	Fly		0.67	0.7768	0.720			
4	一马平川		0.68	0.7553	0.713			
5	Actor-Critic		0.68	0.7393	0.708			
6	AIOT LAB		0.66	0.7382	0.695			
7	BOE		0.62	0.7416	0.676			
8	MCPRL		0.62	0.7224	0.667			
9	shadow		0.61	0.7326	0.665			
10	cv516		0.61	0.7119	0.659			



IMAX
HAPPY HERE
招聘

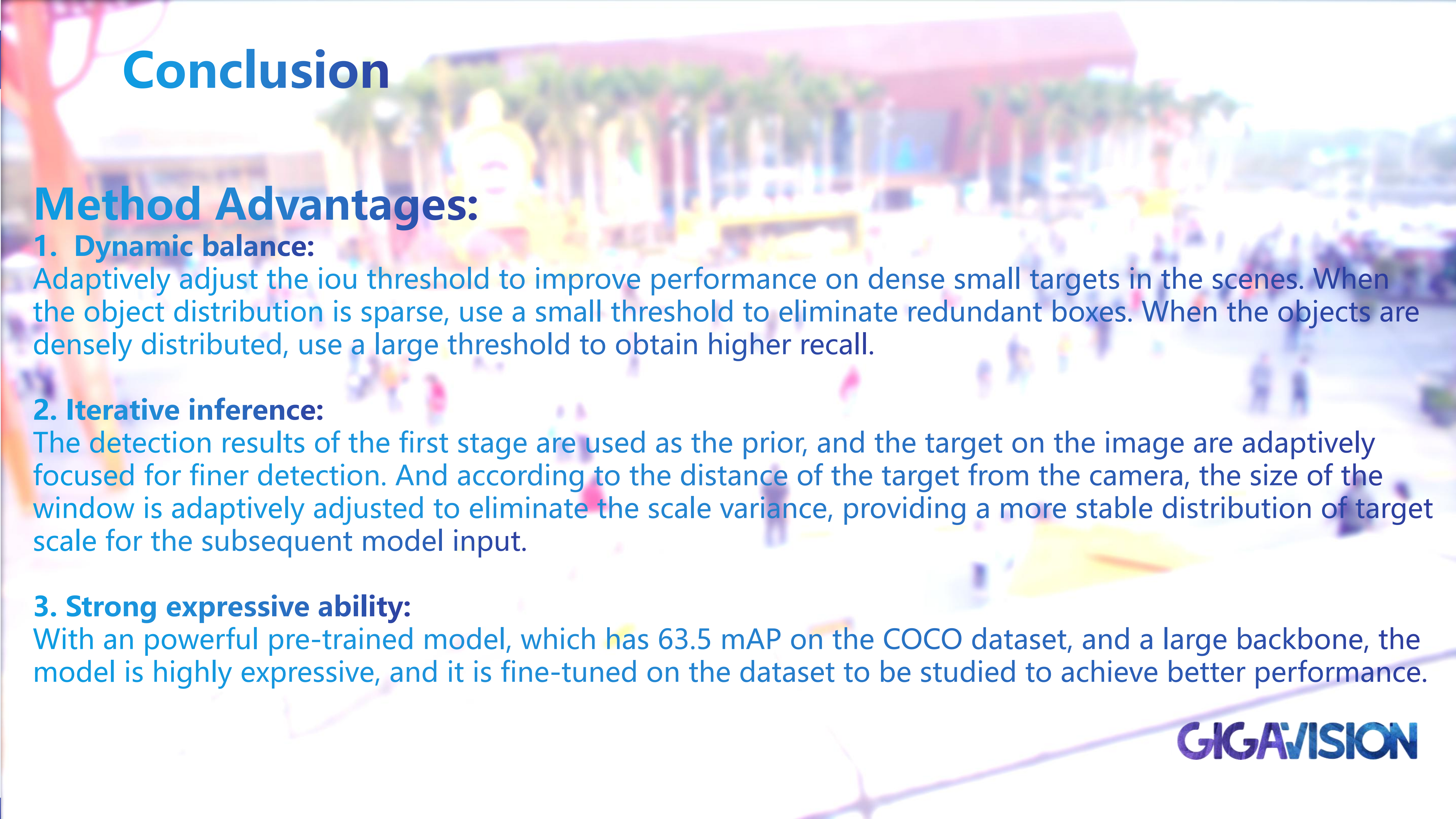

扫黑恶 净环境
促稳定 保平安



扫黑恶 净环境
促稳定 保平安

Catalogue

1. Team Introduction
2. Task Analysis
   a. Scene background complexity
   b. Target distribution variance
3. Solution and Innovation
   a. Data Preprocess
   b. Model Details
   c. Data Postprocess
4. Results & Visualization
5. **Conclusion**



Conclusion

Method Advantages:

1. Dynamic balance:

Adaptively adjust the iou threshold to improve performance on dense small targets in the scenes. When the object distribution is sparse, use a small threshold to eliminate redundant boxes. When the objects are densely distributed, use a large threshold to obtain higher recall.

2. Iterative inference:

The detection results of the first stage are used as the prior, and the target on the image are adaptively focused for finer detection. And according to the distance of the target from the camera, the size of the window is adaptively adjusted to eliminate the scale variance, providing a more stable distribution of target scale for the subsequent model input.

3. Strong expressive ability:

With an powerful pre-trained model, which has 63.5 mAP on the COCO dataset, and a large backbone, the model is highly expressive, and it is fine-tuned on the dataset to be studied to achieve better performance.

Thoughts

THANKS !