NCC 2020

26th National Conference on Communications at IIT Kharagpur

Aerial Multi-Object Tracking by Detection using Deep Association Networks

Ajit Jadhav*, Prerana Mukherjee*, Vinay Kaushik^, Brejesh Lall^

* Dept. of Computer Science, IIIT Sri City, India

^ Dept. of Electrical Engineering, IIT Delhi, India

Introduction

- Tracking-by-detection paradigm
  - Given an input scene and a predefined set of object categories, the task is to locate all the class-level object instances and track the detected candidate boxes in the subsequent frames.
- Applications:
  - Drones are generally used for patroling border areas which cannot be monitored by military forces.
  - The typical application ranges from tracking criminals in surveillance videos, search and rescue operations, sports analysis and scene understanding.

Dataset

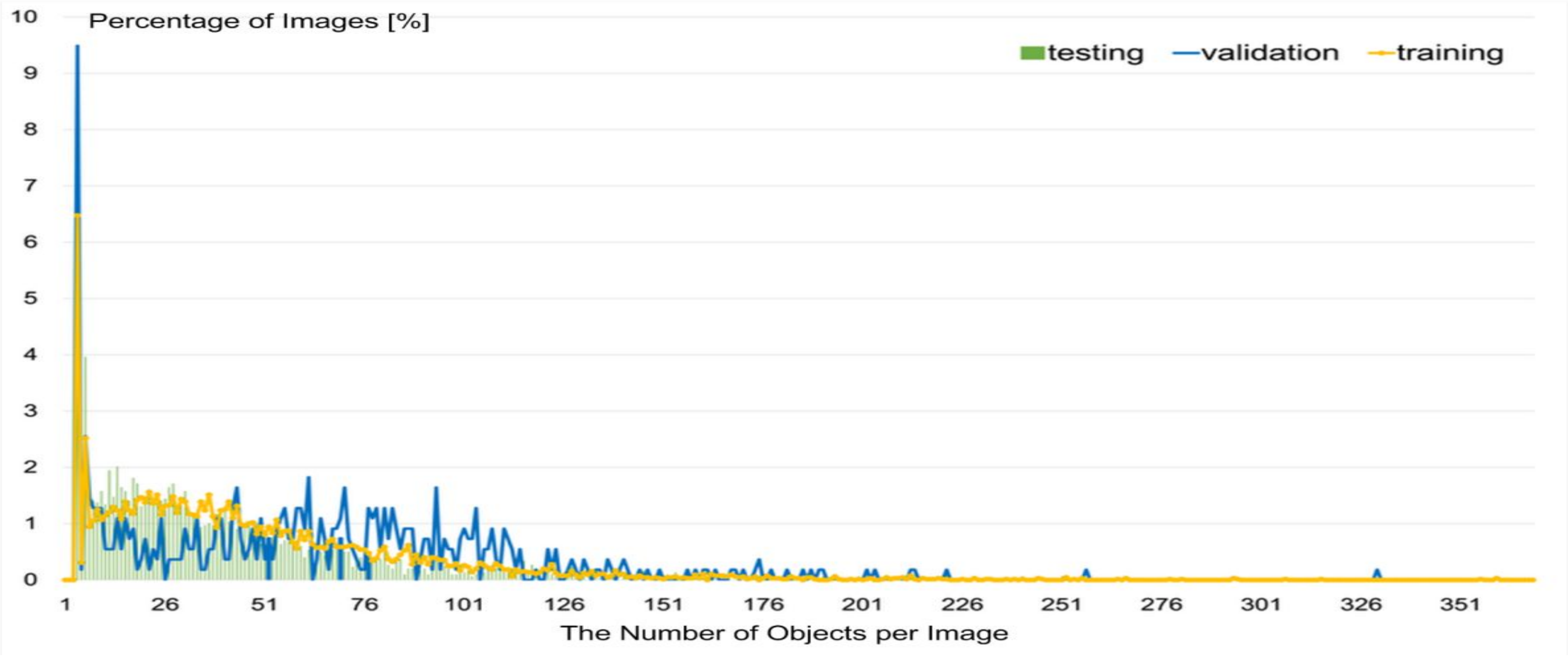
- We use the “Vision meets Drone 2019” i.e. the VisDrone2019 dataset.
- The VisDrone2019 provides a dataset of 10,209 images for this task, with 6,471 images used for training, 548 for validation and 3,190 for testing.
- Contains ten object categories of interest:
  - Pedestrian
  - Person
  - Car
  - Van
  - Bus
  - Truck
  - Motor
  - Bicycle
  - Awning-tricycle
  - Tricycle

Sample images with objects

Key Challenges

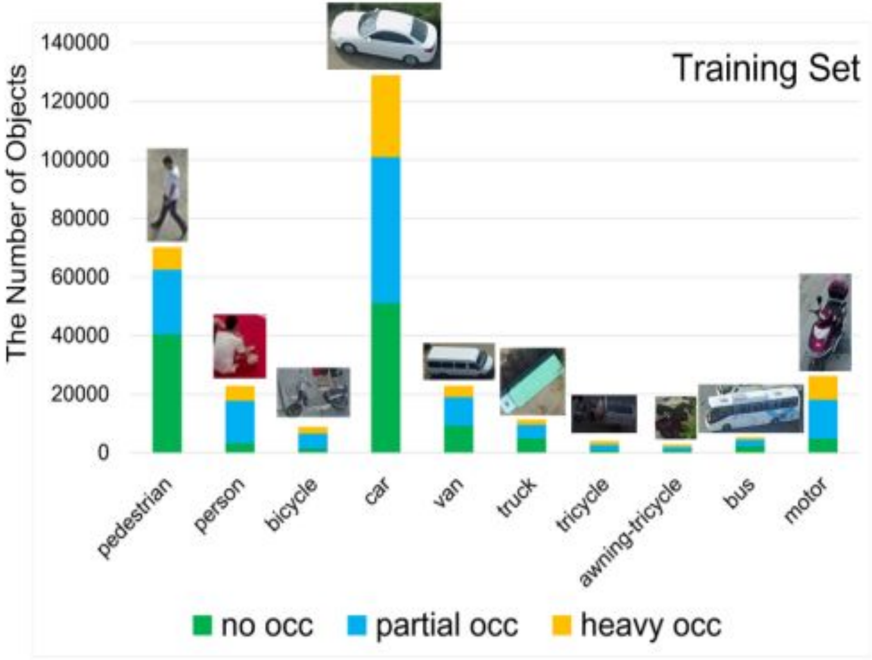
- Dense object distribution
- Large scale variance
- High class imbalance for objects

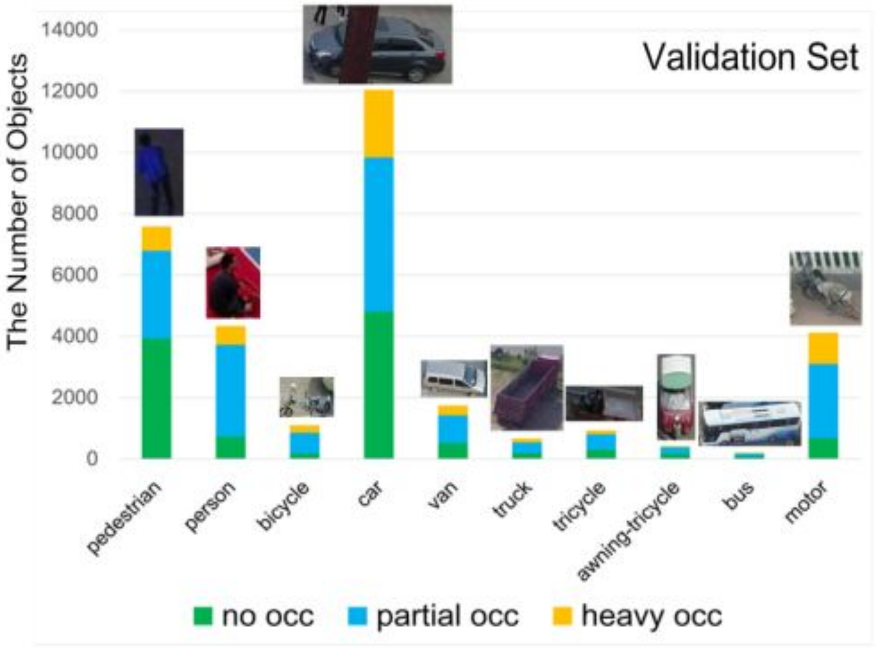
Dense object distribution



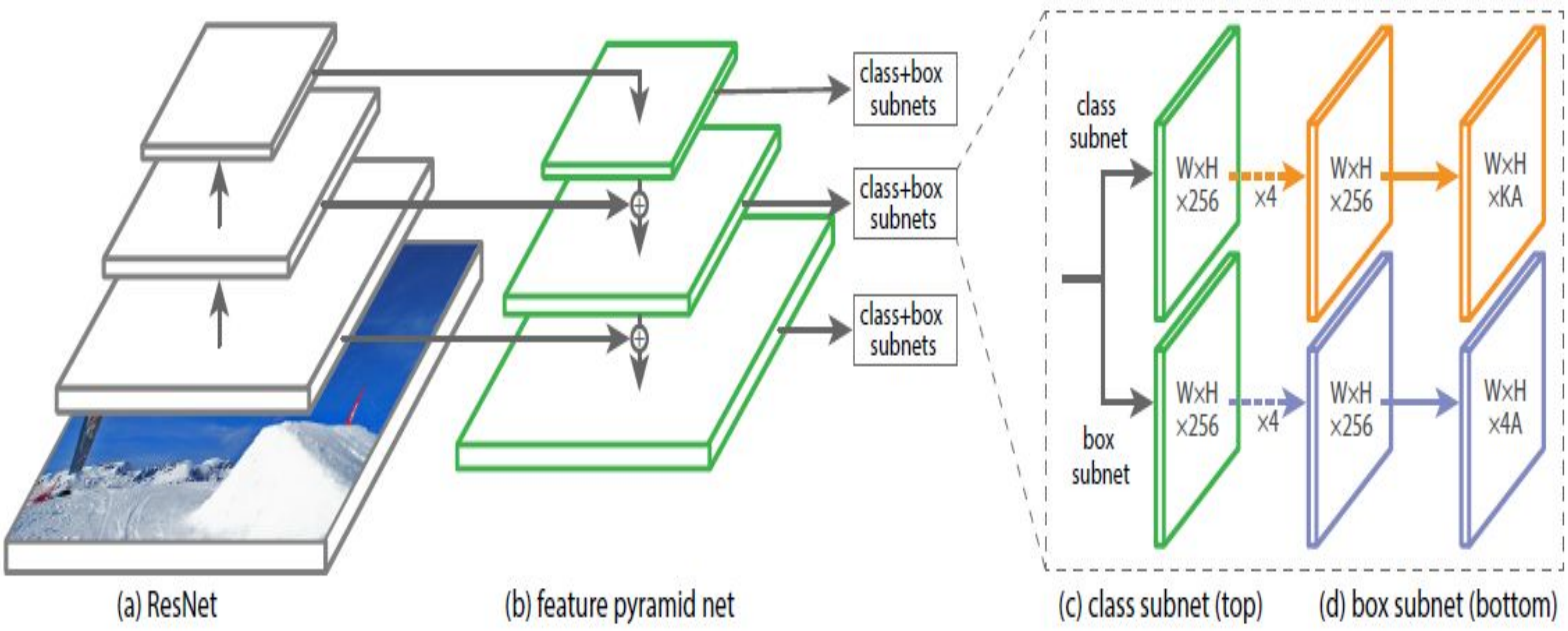
Large Scale Variance

High Class Imbalance





RetinaNet

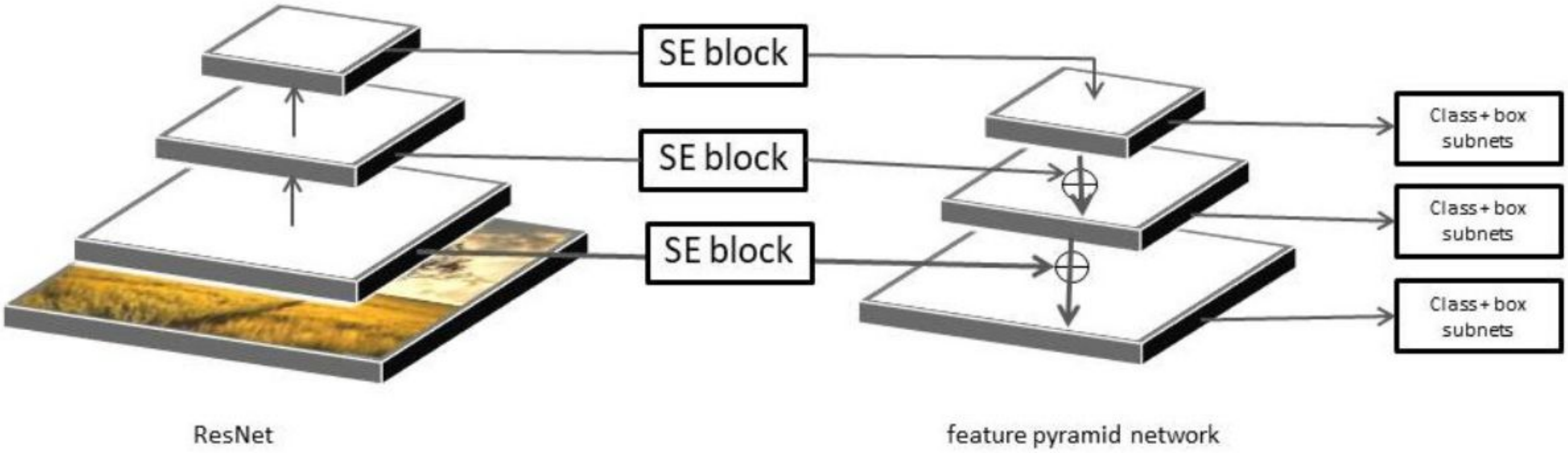


T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. arXiv preprint arXiv:1708.02002, 2017.1, 3, 4

Improvements over RetinaNet:

- Dense scales
- SE attention

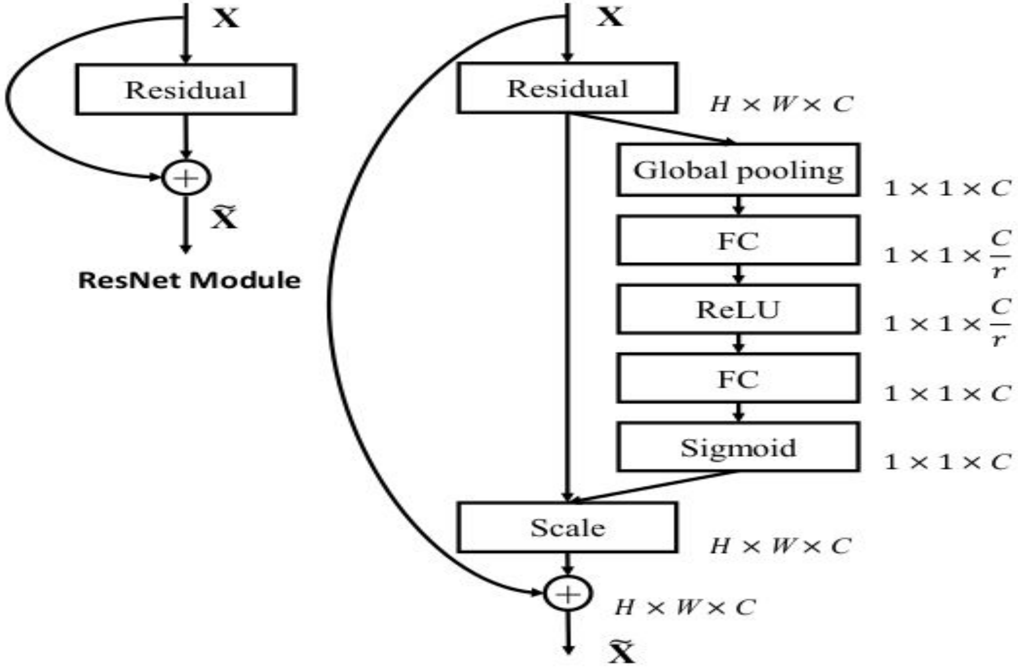
Detection Network



Dense scales

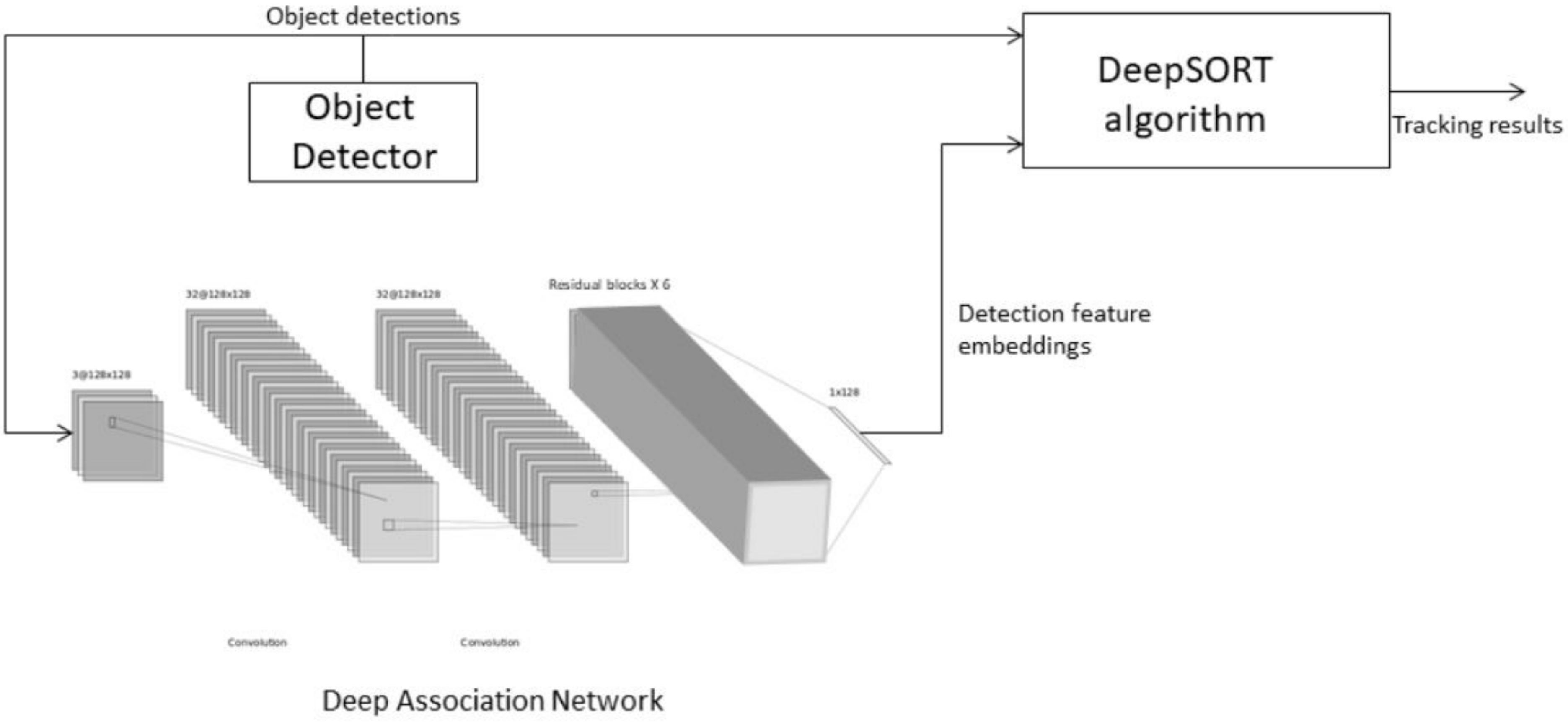
- The anchor parameters used for the original RetinaNet architecture are suited for object detection on natural images.
- To address this issue,we modify the anchor parameters to cover the range of sizes of objects in the dataset.
- Original scales: $\{ 1, 2^{1/3}, 2^{2/3}\}$
- New scales: $\{0.1, 0.25, 0.5, 1, 2^{1/3}, 2.2\}$
- This leads to better detection of objects across all sizes.

SE attention



J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In CVPR, 2018. 1, 2, 3, 4, 7

Tracking Network



Results

Method \AP@IoU	0.50:0.95	0.50	0.75
Yolo v3	13.8	30.43	11.18
RetinaNet	14.45	23.74	15.14
RetinaNet (dense scales)	15.39	33.13	13.07
RetinaNet (dense scales +SE attention)	17.19	37.69	13.97

TABLE I

AVERAGE PRECISION AT MAXDETECTIONS=500

Results

Method \AR@maxDets	1	10	100	500
Yolo v3	0.36	2.63	17.53	19.34
RetinaNet	0.59	5.91	20.96	21.38
RetinaNet (dense scales)	0.48	4.78	22.02	30.49
RetianNet (dense scales +SE attention)	0.52	4.69	23.44	31.93

TABLE II

AVERAGE RECALL AT IOU 0.50:0.95

Results

Method	AP[%]	AP50[%]	AP75[%]	AR1[%]	AR10[%]	AR100[%]	AR500[%]
CornerNet [10]	17.41	34.12	15.78	0.39	3.32	24.37	26.11
Light-RCNN [6]	16.53	32.78	15.13	0.35	3.16	23.09	25.07
DetNet [11]	15.26	29.23	14.34	0.26	2.57	20.87	22.28
RefineDet512 [31]	14.9	28.76	14.08	0.24	2.41	18.13	25.69
Retinanet [12]	11.81	21.37	11.62	0.21	1.21	5.31	19.29
FPN [32]	16.51	32.2	14.91	0.33	3.03	20.72	24.93
Cascade-RCNN [7]	16.09	16.09	15.01	0.28	2.79	21.37	28.43
Ours	11.19	25.65	8.78	0.56	4.87	17.19	24.09

TABLE III
DETECTION RESULTS

Results

Method	AP	AP@0.25	AP@0.50	AP@0.75	AP car	AP bus	AP truck	AP ped	AP van
cem [35]	5.7	9.22	4.89	2.99	6.51	10.58	8.33	0.7	2.38
cmot [36]	14.22	22.11	14.58	5.98	27.72	17.95	7.79	9.95	7.71
gog [37]	6.16	11.03	5.3	2.14	17.05	1.8	5.67	3.7	2.55
h2t [38]	4.93	8.93	4.73	1.12	12.9	5.99	2.27	2.18	1.29
ihtls [39]	4.72	8.6	4.34	1.22	12.07	2.38	5.82	1.94	1.4
Ours	13.88	23.19	12.81	5.64	32.2	8.83	6.61	18.61	3.16

TABLE IV
TRACKING RESULTS

Conclusion

- Dense anchor scales with large scale variance correctly detect the dense distribution of smaller objects.
- Squeeze-and-Excitation (SE) blocks capture the channel dependencies resulting in better feature representation for the detection task in moving camera constraints.
- Training deep association network on the object hypotheses generated from the detection module and feeding the same to the the deep sort algorithm leads to better tracking.
- Large number of average confidence detections are preferable than less number of high confidence detections to build an optimal tracker.
- The tracking can be further improved by better data augmentation methods, collecting more relevant data and incorporating structure similarity losses.

References

- T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. arXiv preprint arXiv:1708.02002, 2017.1, 3, 4
- J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In CVPR, 2018. 1, 2, 3, 4, 7
- P. Zhu, L. Wen, D. Du, X. Bian, H. Ling, Q. Hu, Q. Nie, H. Cheng, C. Liu, X. Liu et al., “Visdrone-det2019: The vision meets drone object detection in image challenge results,” in Proceedings of the International Conference on Computer Vision (ICCV), 2019, pp. 0–0.
- N. Wojke and A. Bewley, “Deep cosine metric learning for person re-identification,” in 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018, pp. 748–756.
- N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” in 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017, pp. 3645–3649.

Thank you