RRNet 复现

朱冠铭

摘要

无人机和监控摄像头在城市场景中捕捉到的物体通常大小不一，密度极高。因此，在 [1] 中提出了一种称为 RRNet[1]的混合检测器，用于此类挑战性任务中的目标检测。我们将无锚检测器与重回归模块混合以重新构建检测器。先前锚的丢弃使我们的模型摆脱了边界框大小回归的困难任务，从而在密集场景中实现了更好的多尺度目标检测性能。基于无锚的检测器首先生成粗框。然后，在粗略预测上应用位置敏感的感兴趣区域对齐模块（Position Sensitive ROI-Align）以产生精确的边界框。此外，在 RRNet[1]原有的自适应重采样数据增强策略上，我们加入了更加高精度的背景分割数据集。我们的实验表明，改进后的 RRNet 在 VisDrone2019 数据集上优于原来的 RRNet 数据。

关键词：位置敏感；自适应采样；背景分割。

1 引言

无人机已经在学术界和工业界中取得了广泛的应用。它要求我们理解并分析它们捕获的图像数据。在深度学习时代，基于 DNN（深度神经网络）的目标检测器[2-3]显著提高了目标检测的性能。然而，正常拍摄的图像与无人的拍摄图像之间存在许多显著差异；这些差异使得物体检测成为一项具有挑战性的任务。首先，这些图像中的对象具有不同的比例。如图 1 所示，远的物体非常小，近的物体很大。此外，城市中有许多密集的场景（如图 1）。密集度会导致大量遮挡，使得物体检测更加困难。

图 1: 城市密集场景

通常，当前基于深度学习的对象检测器分为两类。第一种是两级检测器[4]。他们使用区域提议网络来确定先前的锚是目标还是背景。先前的锚是几个手动定义的潜在边界框。然后，他们使用两个头部网络将潜在锚点分类为一组类别，并估计锚点和地面真相框之间的偏移。另一类被称为单级检测器[5]。与两级检测器不同，一级检测器丢弃了区域建议网络。他们直接使用两个检测器来预测先前锚

的类别和偏移。在低分辨率图像网格上生成这两种类型检测器的先前锚点。根据 IoU（联合上的交点），每个先前锚点只能指定一个对象边界框。然而，在无人机拍摄的图像中，固定形状的锚很难处理各种尺度的物体。最近，提出了另一种类型的检测器，即无锚检测器。它们将边界框预测减少到关键点和大小估计。它提供了一种更好的方法来检测不同尺度的物体。尽管如此，大小上的巨大差异（例如，从 10 到 1000）会使得回归更加困难。在文中，作者提出了一种称为 RRNet 的混合检测器。尽管物体存在不同尺度，但是物体的中心点总是存在的。因此，我们使用两个检测器来预测每个对象的中心点以及宽度和高度，而不是使用锚框然后，我们将这些中心点和大小转换为粗略的边界框。最后，我们将深度特征图和粗糙边界框输入到重新回归模块中。重新回归模块可以调整粗略边界框并生成最终精确边界框。此外，一些证据[6]表明，良好的数据扩充甚至可以在不改变网络架构的情况下提升深度模型，以实现最先进的性能。因此，原文提出了一种称为自适应重采样（AdaRessampling）的数据增强策略。该策略可以在逻辑上增强图像上的对象。

2 相关工作

2.1 数据增强

数据增强的目的是消除训练数据集和测试数据集之间的偏差。深度模型通常使用许多数据增强，例如随机裁剪和随机翻转，以避免过度拟合。Zoph 等人[6]使用自动机器学习（AutoML）来搜索最佳增强策略。他们在不改变任何网络架构的情况下实现了最先进的技术。Kisantal 等人[7]使用复制粘贴来提高小对象的性能。他们首先使用分割遮罩裁剪小对象，然后在图像中随机粘贴裁剪的小对象。然而，我们不能简单地将裁剪的对象随机粘贴到无人机拍摄的图像中。我们注意到，在无人机拍摄的图像中有一个明显的位置在先。例如，汽车在空中飞行是不可能的。因此，RRNet 提出了一种全新的自适应数据增强方法 AdaResample。

2.2 基于锚点的对象检测

锚点（anchor）方法被大多数现有的探测器广泛采用。两级检测器长期以来一直是目标检测领域的主要方法。Fater-RCNN[2]提出了区域提案网络（RPN）来生成 Proposal。然后，Proposal 被发送到第二阶段以生成最终的边界框。大多数其他两阶段方法[8]是 Faster-RCNN 的变体。此外，还提出了一些多级检测器。级联 RCNN[4]扩展了更快的 RCNN[2]，以解决过度拟合和质量不匹配的问题。与两阶段和多阶段方法相比，单阶段方法没有建议生成阶段，并在一个部分中预测边界框。虽然它们不会产生概率，但单级方法仍然使用锚盒。SSD[5]和 YOLO[9]直接对一个 chors 进行分类和回归，以获得最终的边界框。RetinaNet[10]引入焦点损失，通过重塑标准交叉熵损失来解决类别失衡问题。

2.3 无锚物体检测

最近，一些检测器丢弃了先前的锚方法。他们将目标检测任务转换为关键点和大小估计。Corner-Net[11]检测边界框角作为关键点，然后在后期处理中匹配左上角和右下角，而 CenterNet[3]简单地导出每个对象的单个中心点，并预测其宽度和高度。FoveaBox[12]预测对象现有可能性的类别敏感语义图，并为可能包含对象的每个位置生成类别不可知的边界框。

3 本文方法

3.1 本文方法概述

在 RRNet 主要由自适应数据增强和两部分组成。图 2的顶部是 RRNet 的架构。我们首先将图像馈送到一些卷积块中以获得初始特征图。之后，两个 HourGlass 块（HGBlock）[13]提取具有更多语义信息的鲁棒特征图。我们将这些特征输入到两个独立的检测器中。heatmap 检测器产生对象中心点的类别敏感概率热图。此外，另一个检测器将给出所有中心点的尺寸估计。紧接着将热图和尺寸预测转换为粗略的边界框。最后，将重新回归模块应用于这些粗检测框以生成细化的边界框。

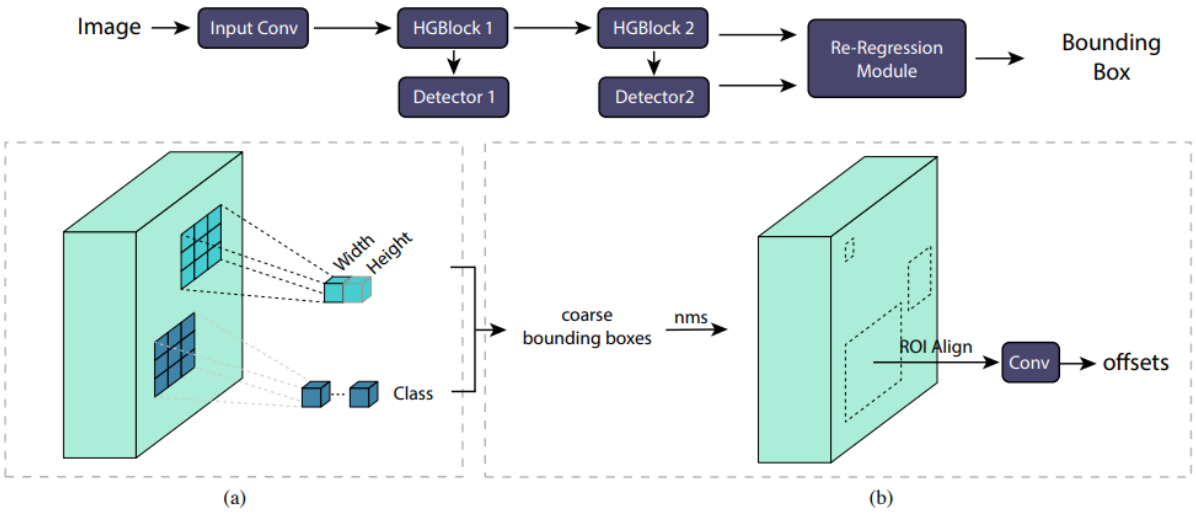

图 2: RRNet 的架构。（a）RRNet 中的检测器。篇幅受限只提供一个特征像素的预测。实际算法会对所有特征像素执行相同的预测。（b）重回归模块。

3.2 重回归模块 Re-Regression Net

重回归模块（Re-Regression）对我们的模型进行细化粗略的边界框的操作。我们将 HGBlock2 生成的特征图和粗略边界框输入到重回归模块中。重回归模块与 Faster-RCNN 头部相似，但不包括分类网络。RRNet 首先使用 NMS 算法来过滤重复的边界框。之后，RRNet 使用 ROI 对齐来对齐特征，并使用两个卷积层来预测偏移值。最后，RRNet 将偏移值应用于粗略边界框以获得最终预测。

3.3 自适应重采样模块 AdaResampling

为了消除这两种失配，RRNet 中提出了一种称为 AdaResampling 的自适应增强策略。图 3显示了 AdaRes 采样的流程图。一开始，算法将无人机拍摄的图像输入到预训练的语义分割网络中，以获得先前的分割背景图。由于无人机捕获图像与用于分割网络训练的数据集之间的差异，分割网络可能会产生噪声结果。我们不需要很高的召回值，但需要对道路进行高度的预测。因此，我们使用侵蚀算法和 3×3 中值滤波器来尽可能地去除假道路区域。然后，我们根据路线图对有效位置进行采样，以放置增强对象。之后，通过尺度变换函数调整裁剪对象的大小。高宽比是恒定的。

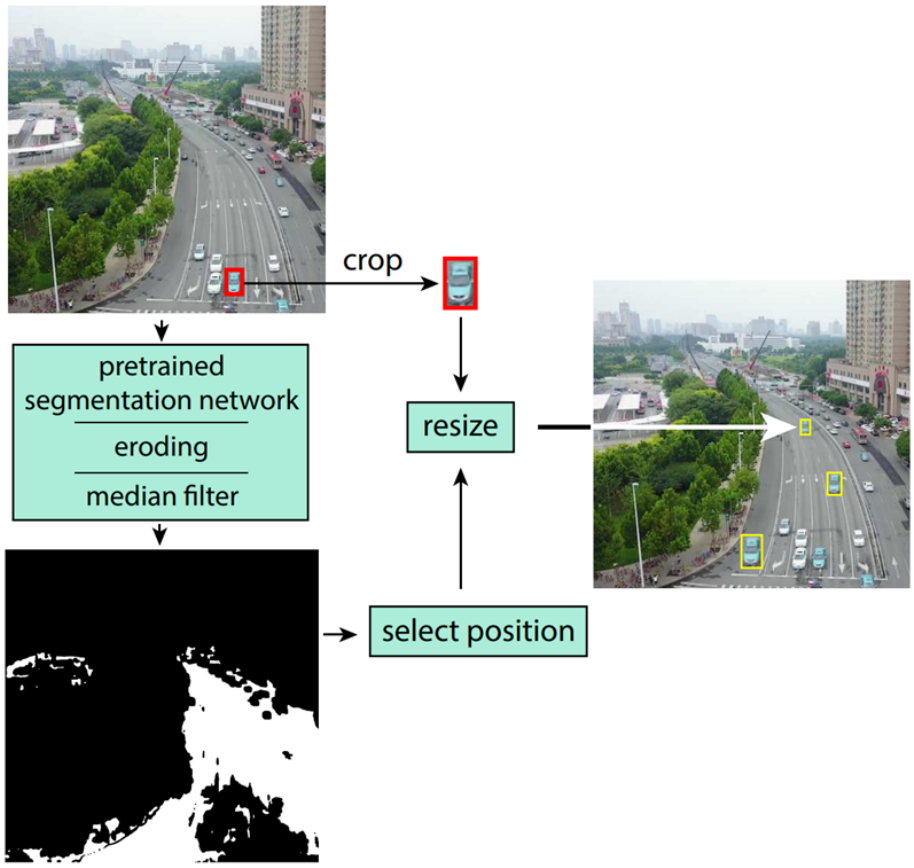

图 3: AdaResampling 的流程图

图 3的右侧部分是我们的 AdaResampling 生成的训练图像。我们可以看到，汽车只能放置在道路上，并且增强对象的比例是合适的。

4 复现细节

4.1 与已有开源代码对比

RRNet 的代码已经开源，本文工作主要在此基础上进行改进。主要优化数据集中较多的小轿车和路上行人。

4.2 实验环境搭建

与大多数深度神经网络类似，同样使用水平翻转和随机裁剪作为简单的数据增强。训练阶段的图像大小为 512×512。使用前面提到的的 AdaResampling 来增加人员、行人、自行车、三轮车、遮阳篷三轮车和电动车的样本数量。密度系数 d 同样设置为 0.00005。在我们的实验中，我们采用 Adam 作为优化器。每个小批量每 GPU 有 4 个图像，我们在 2 个 GPU 上训练我们的模型进行 100k 次迭代，学习率为 2.5e-4，在 60k 和 80k 次迭代时降低了 10。用于分类的损失函数是 focal loss。

4.3 创新点

1. 更换分割效果更好的数据集根据前面的分析，RRNet 的高性能来源于重采样的数据增强，而道路的分割准确率会影响到车辆的识别效率，所以本文选择了 2022 年的 SOTA 方法 ViT-Adapter-L[14]来得到分割后的 road 数据集。取消过程中使用中值滤波器，对比实验发现直接使用分割得到的 mask 最终准确率比使用滤波器平滑边缘的准确率要高 1 2

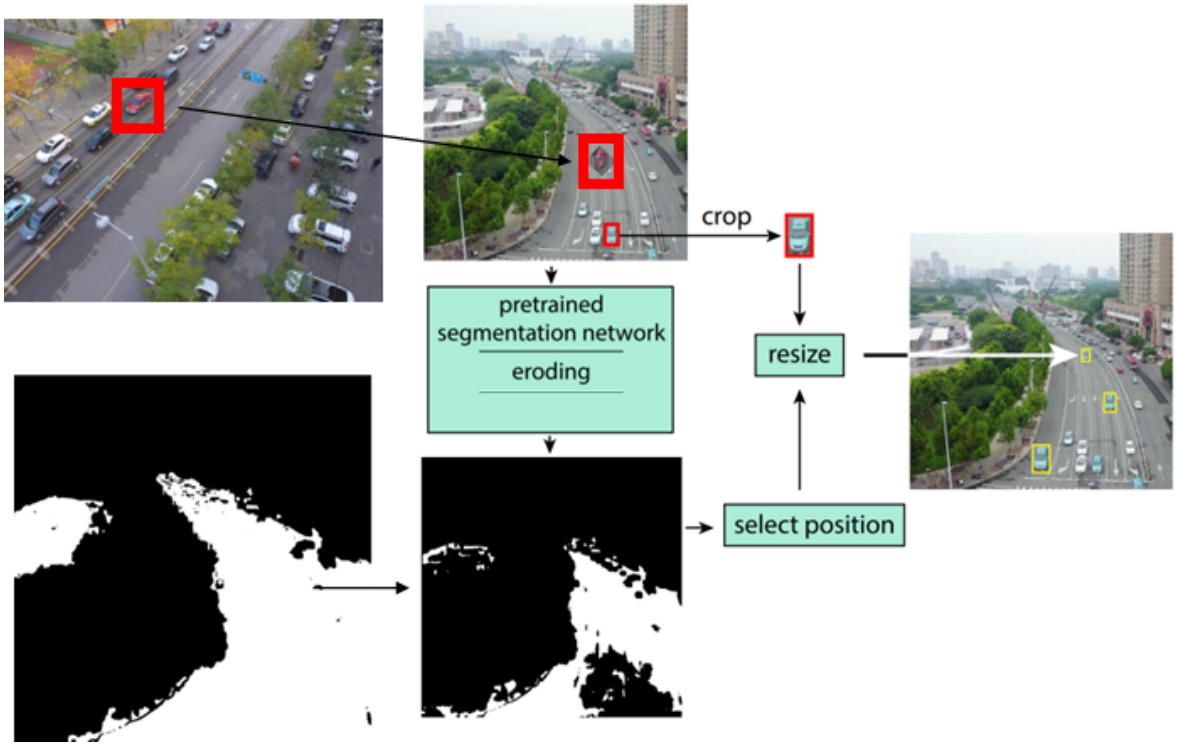

图 4: 改进后的 AdaResampling 的流程图，可以看到改进后的道路分割更加准确。

2. 区域敏感 ROI 对齐粗检测网络后面紧接着 Re-Regression 模块，生成精细化的边界框。Re-Regression 模块可以细化我们的模型粗边界框。我们将 HGBlock 2 生成的特征图和粗边界框输入 Re-Regression 模块。Re-Regression 模块类似于 Faster-RCNN 头部，但不包括分类网络。首先使用 NMS 算法对候选边界框进行过滤。在此之后，我们使用 ROI -Align 来对齐特征，并使用两个卷积层来预测偏移值。最后，将偏移值应用到粗边界框中，得到最终的预测结果。

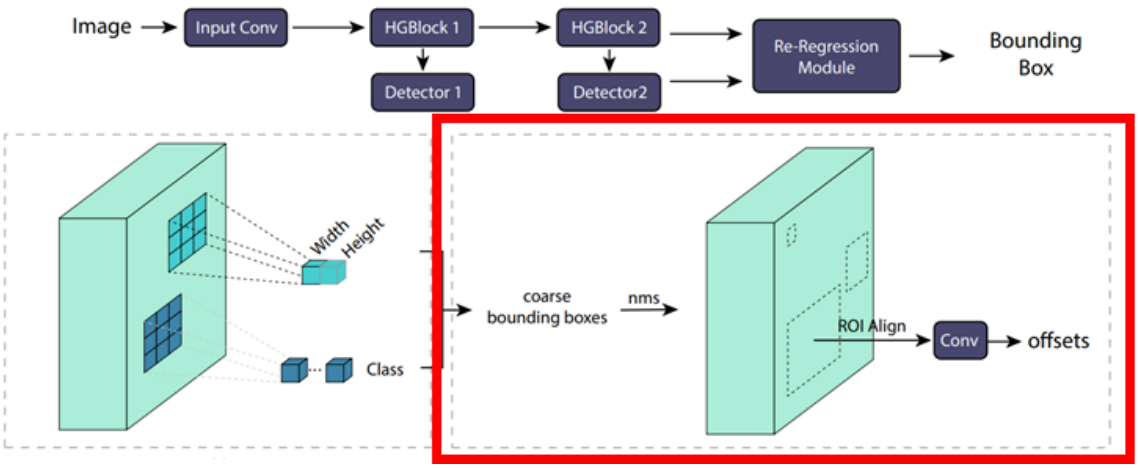

图 5: 进行改进的 Re-Regression 模块

目标检测中的经典网络如 Fast R□CNN 等大多只是利用了深度神经网络的最后层来进行预测，然而由于空间和细节特征信息的丢失，难以在深层特征图中检测小目标。RRNet 中使用的 ROI Align 是在 Mask-RCNN 里提出的一种区域特征聚集方式, 很好地解决了 ROI Pooling 操作中两次量化（取整）造成的区域不匹配 (mis-alignment) 的问题，小目标由于不匹配会导致 1 5 像素的误差，这个无疑是会大大降低准确率的。实验显示，在检测测任务中将 ROI Pooling 替换为 ROI Align 可以提升检测模型的准确性。PsRoIAlign 可以保存局部区域的位置敏感性，这是考虑数据集中较大量的小目标对象而进行的优化。这里需要引入位置信息，借鉴使用了 Light-Head R-CNN[15]的实现。

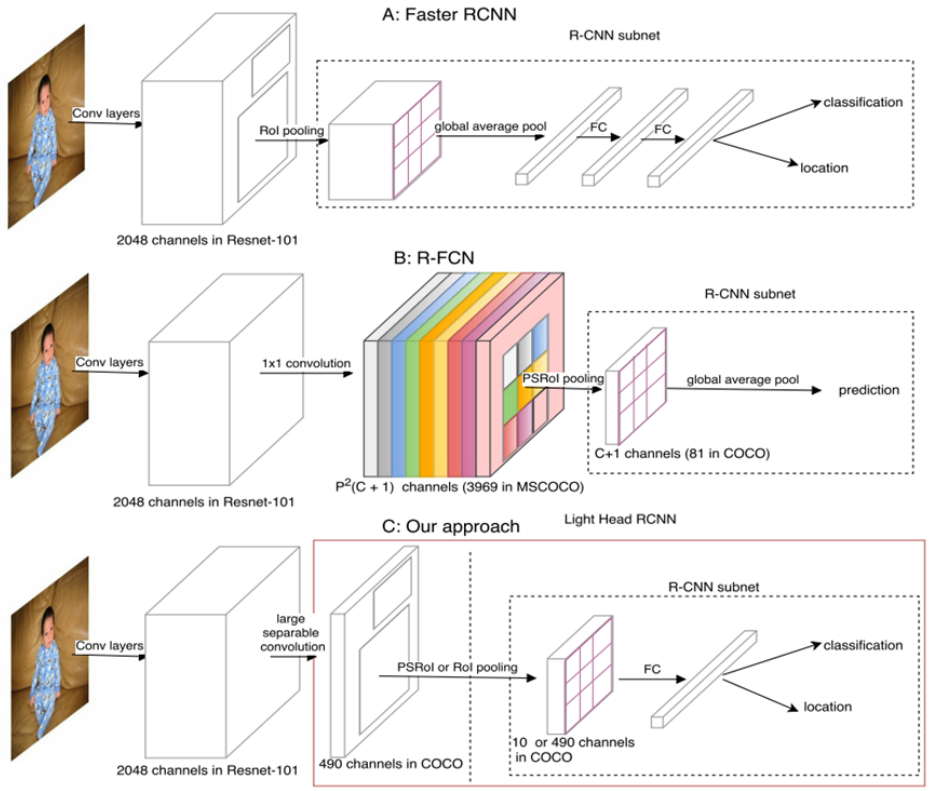

图 6: PSRoI-Align 结构图

5 实验结果分析

本部分对实验所得结果进行分析，对实验内容进行说明，实验结果进行描述并分析。使用 RRNet 同样的评价脚本得出了如图 7 的最终效果。

```
cheny@server:/home/data2/RRnet 2023-01-18 13:22:56
$ conda activate zgm
(zgm) python eval.py
Start generating Txt file ...
⇒ Use GPU: 0
⇒ Use GPU: 1
[266/274]⇒ Evaluation Done!
[274/274]⇒ Evaluation Done!
Start Evaluating ...
Average Precision  (AP) @[ IoU=0.50:0.95] = 0.3422.
Average Precision  (AP) @[ IoU=0.50      ] = 0.5418.
Average Precision  (AP) @[ IoU=0.75      ] = 0.3669.
Average Recall     (AR) @[ IoU=0.50:0.95] = 0.446.
Cost Time: 13.352316141128545s
```

图 7: 实验结果示意

在 VisDrone2018 数据集上，RRNet 实现了 0.3241 mAP 的可靠结果。RR 模块生成的梯度对于主干和检测器优化是有利的。我们改进后的 mAP 比原先的算法提升约 2%，虽然在 AP50 的准确率有所下降，但是 AP75 的准确率有较大提升。这主要得益于预训练过程中生成了较准确的在路目标和区域敏感对齐。

Methods	mAP	AP50	AP75	AR1	AR10	AR100	AR500
RetinaNet [12]	11.81	21.37	11.62	0.21	1.21	5.31	19.29
RefineDet [20]	14.90	28.76	14.08	0.24	2.41	18.13	25.69
DetNet [19]	15.26	29.23	14.34	0.26	2.57	20.87	22.28
Cascade RCNN [2]	16.09	16.09	15.01	0.28	2.79	21.37	28.43
CornerNet [9]	17.41	34.12	15.78	0.39	3.32	24.37	26.11
FPN [11]	16.51	32.20	14.91	0.33	3.03	20.72	24.93
Light-RCNN [10]	16.53	32.78	15.13	0.35	3.16	23.09	25.07
ACM-OD†	29.13	54.07	27.38	0.32	1.48	9.46	44.53
DPNet-ensemble†	**29.62**	54.00	**28.70**	0.58	3.69	17.10	42.37
RRNet	29.13	**55.82**	27.23	**1.02**	**8.50**	**35.19**	**46.05**

Category	ped	people	bicycle	car	van	truck	tricycle	awn	bus	motor
mAP	30.442	14.851	13.724	51.427	36.143	35.224	28.019	18.999	44.204	25.854

图 8: VisDrone2018 测试集的性能。† 标记的是 ICCV VisDrone2019 图像中物体检测挑战赛的冠军和季军。RRNet 是亚军。

对比目前 SOTA 的算法，准确率还有不小提升改进的空间。

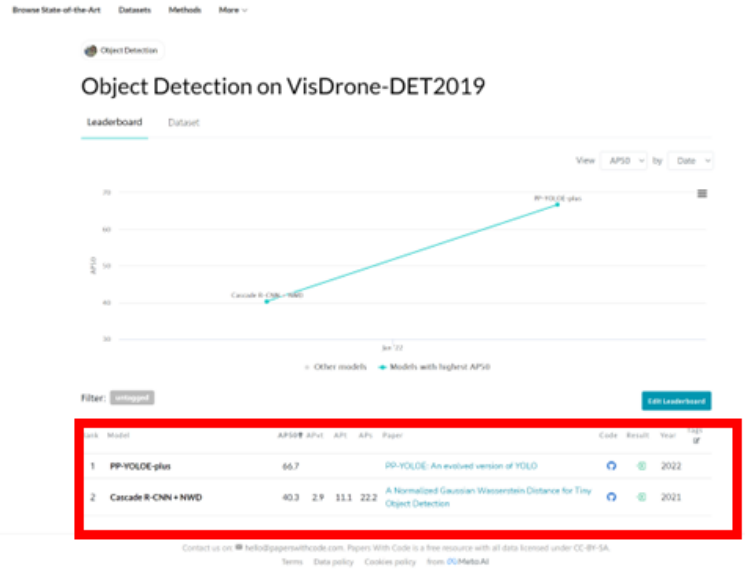

图 9: 截至目前该数据集的 SOTA 算法

6 总结与展望

在本文中，我们提供了改进 RRNet 的一些新技巧思路并取得成果。它在密集场景中的非常小的对象上表现出优异的性能。但是本文的实验动机需要进一步验证和补充，同时在消融实验方面，需要对每一加入模块的有效性进一步验证。

参考文献

[1] ZHU P, WEN L, BIAN X, et al. RRNet: A Hybrid Detector for Object Detection in Drone-captured Images.[J]. The VisDrone 2019，Computer Vision and Pattern Recognition, 2019.

[2] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.

[3] ZHOU X, WANG D, KRAHENBUHL. P. Objects as Points.[J]. arXiv, 2018.

[4] CAI Z, VASCONCELOS. N. Cascade R-CNN: Delving into High Quality Object Detection.[J]. Computer Vision and Pattern Recognition, 2018, 08(08): 6154-6162.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector.[J]. European Conference on Computer Vision (ECCV)., 2019: 21-37.

[6] ZOPH B, CUBUK E D, GHIASI G, et al. Learning Data Augmentation Strategies for Object Detection.[J]. arXiv., 2019.

[7] KISANTAL M, WOJNA Z, MURAWSKI J, et al. Augmentation for small object detection.[J]. arXiv., 2019.

[8] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature Pyramid Networks for Object Detection.[J]. arXiv., 2019.

[9] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection.[J]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)., 2016: 779-788.

[10] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection.[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence., 2018, PP(99): 1-1.

[11] LAW H, DENG. J. CornerNet: Detecting Objects as Paired Keypoints.[J]. European Conference on Computer Vision (ECCV)., 2018: 765-781.

[12] KONG T, SUN F, LIU H, et al. FoveaBox: Beyond Anchor-based Object Detector.[J]. arXiv, 2019.

[13] NEWELL A, YANG K, DENG. J. Stacked Hourglass Networks for Human Pose Estimation.[J]. European Conference on Computer Vision (ECCV), 2016: 483-499.

[14] CHEN Z, DUAN Y, WANG W, et al. Vision Transformer Adapter for Dense Predictions.[J]. Computer Vision and Pattern Recognition (CVPR), 2022.

[15] LI Z, PENG C, YU G, et al. Light-Head R-CNN: In Defense of Two-Stage Object Detector.[J]. Computer Vision and Pattern Recognition (CVPR), 2017.