计算机视觉课程报告

学号：　10185102144

姓名：　董辰尧

专业名称：	计算机科学与技术
学生年级：	2018 级本科生
课程性质：	专业选修
研修时间：	2020～2021 学年第 2 学期

计算机科学与技术学院

2021 年 6 月

课程内容统计

- 请自评你的项目完成情况，在表中相应位置划√。

课程学习自我评价

内容\评价	阅读文献 0—5 篇	阅读文献 5—10 篇	阅读文献 10 篇以上	代码 实现
第 12 章 目标识别	√			

总体课程学习情况自我评价

完成情况	尚未 完成	基本 完成	较好 完成	圆满 完成
总体情况		√		

第 12 章 目标识别

一、这一章学习中你的工作

二、查阅文献清单

格式：

[1] Wang G , Luo C , Sun X , et al. Tracking by Instance Detection: A Meta-Learning Approach[J]. IEEE, 2020

[2]　F　Yu,　Li W ,　Li Q , et al. POI: Multiple Object Tracking with High Performance Detection and Appearance Feature[C]// European Conference on Computer Vision. Springer, Cham, 2016..

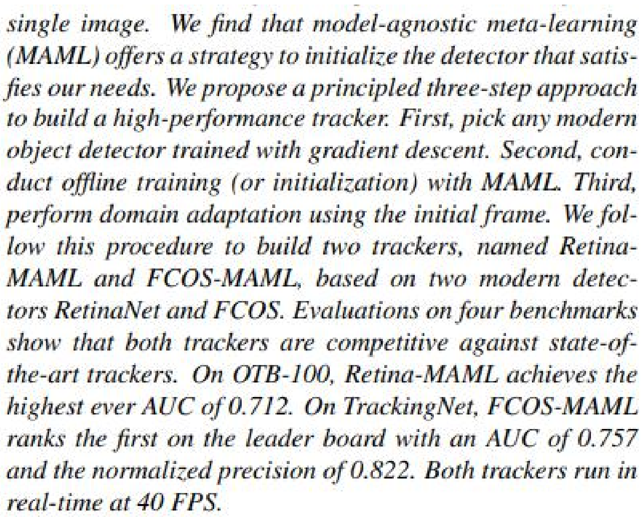
[3]　Zhang W ,　Zhou H ,　Sun S , et al. Robust Multi-Modality Multi-Object Tracking[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019.

三、文献解读

1. 文献 1

（a）文献名： J Wang, G. , Luo, C. , Sun, X. , Xiong, Z. , & Zeng, W. . (2020). Tracking by instance detection: a meta-learning approach. IEEE.

（b）主要创新思想



single image. We find that model-agnostic meta-learning (MAML) offers a strategy to initialize the detector that satisfies our needs. We propose a principled three-step approach to build a high-performance tracker. First, pick any modern object detector trained with gradient descent. Second, conduct offline training (or initialization) with MAML. Third, perform domain adaptation using the initial frame. We follow this procedure to build two trackers, named Retina-MAML and FCOS-MAML, based on two modern detectors RetinaNet and FCOS. Evaluations on four benchmarks show that both trackers are competitive against state-of-the-art trackers. On OTB-100, Retina-MAML achieves the highest ever AUC of 0.712. On TrackingNet, FCOS-MAML ranks the first on the leader board with an AUC of 0.757 and the normalized precision of 0.822. Both trackers run in real-time at 40 FPS.

这篇论文的动机是通过元学习(Meta-Learning)直接将一个目标检测器转化为一个 high-performance 的跟踪器。

这篇论文提出通过三步来构建一个 high-performance 的跟踪器:

1.选择一个在目标检测任务中表现好的检测器.

2.通过元学习算法 MAML 对该检测器进行离线训练.

3.通过首帧图像对该模型进行 domain adaptation.

（c）主要原理剖析及说明

第一步，是选择一个比较好的检测模型，这篇论文中选择了 RetinaNet 和 FCOS 两个检测模型。

第二步，根据元学习的思想，将当前训练集分为 support set(即 train set)和 target set(即 test set)，然后通过 MAML 算法对模型进行离线训练（如下图所示）

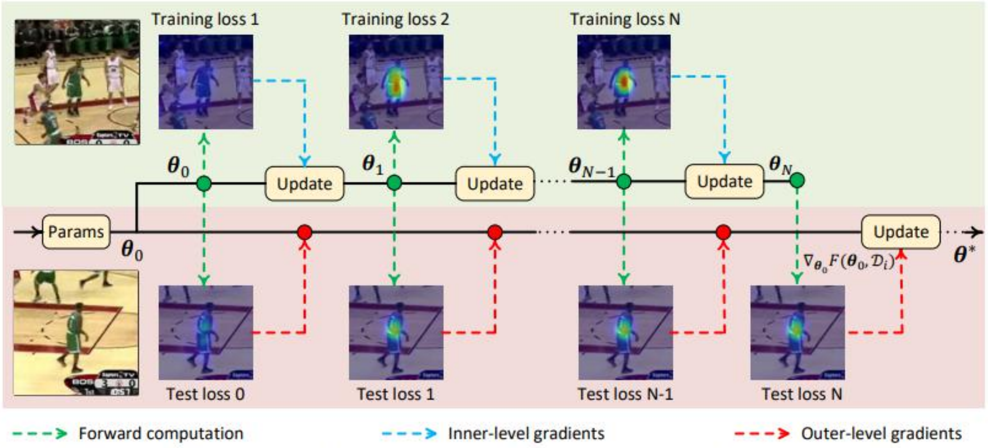
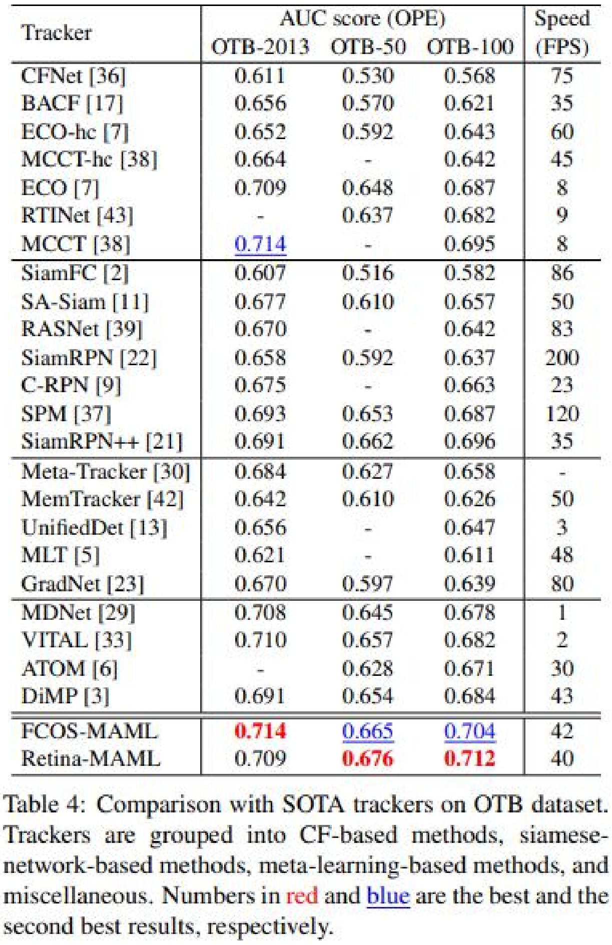


Figure 2: Illustration of our training pipeline. The first row is the inner training loop. A few steps of SGD optimization is performed on the support images. The updated parameters after each step are used for calculating the meta-gradient based on testing images. Best viewed in colors.

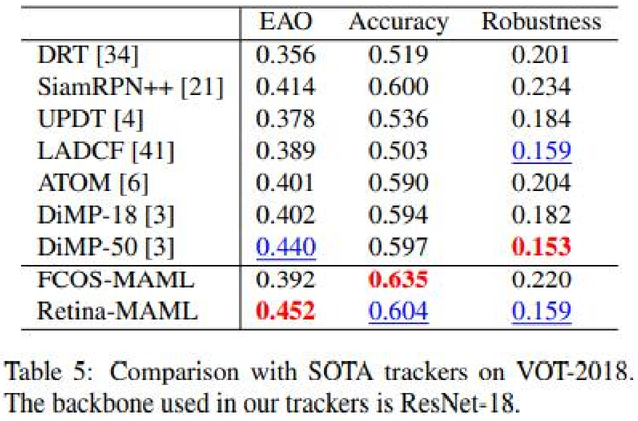
第三步，在一个新的 video sequence 中进行 domain adaptation,即通过对首帧图像进行图像增广得到一组 support set,之后通过梯度下降法对模型进行一轮参数更新

（d）主要实验结果（现有原文章中的）



Tracker	AUC score (OPE)			Speed
	OTB-2013	OTB-50	OTB-100	(FPS)
CFNet [36]	0.611	0.530	0.568	75
BACF [17]	0.656	0.570	0.621	35
ECO-hc [7]	0.652	0.592	0.643	60
MCCT-hc [38]	0.664	-	0.642	45
ECO [7]	0.709	0.648	0.687	8
RTINet [43]	-	0.637	0.682	9
MCCT [38]	0.714	-	0.695	8
SiamFC [2]	0.607	0.516	0.582	86
SA-Siam [11]	0.677	0.610	0.657	50
RASNet [39]	0.670	-	0.642	83
SiamRPN [22]	0.658	0.592	0.637	200
C-RPN [9]	0.675	-	0.663	23
SPM [37]	0.693	0.653	0.687	120
SiamRPN++ [21]	0.691	0.662	0.696	35
Meta-Tracker [30]	0.684	0.627	0.658	-
MemTracker [42]	0.642	0.610	0.626	50
UnifiedDet [13]	0.656	-	0.647	3
MLT [5]	0.621	-	0.611	48
GradNet [23]	0.670	0.597	0.639	80
MDNet [29]	0.708	0.645	0.678	1
VITAL [33]	0.710	0.657	0.682	2
ATOM [6]	-	0.628	0.671	30
DiMP [3]	0.691	0.654	0.684	43
FCOS-MAML	**0.714**	0.665	0.704	42
Retina-MAML	0.709	**0.676**	**0.712**	40

Table 4: Comparison with SOTA trackers on OTB dataset. Trackers are grouped into CF-based methods, siamese-network-based methods, meta-learning-based methods, and miscellaneous. Numbers in red and blue are the best and the second best results, respectively.



	EAO	Accuracy	Robustness
DRT [34]	0.356	0.519	0.201
SiamRPN++ [21]	0.414	0.600	0.234
UPDT [4]	0.378	0.536	0.184
LADCF [41]	0.389	0.503	0.159
ATOM [6]	0.401	0.590	0.204
DiMP-18 [3]	0.402	0.594	0.182
DiMP-50 [3]	0.440	0.597	**0.153**
FCOS-MAML	0.392	**0.635**	0.220
Retina-MAML	**0.452**	0.604	0.159

Table 5: Comparison with SOTA trackers on VOT-2018. The backbone used in our trackers is ResNet-18.

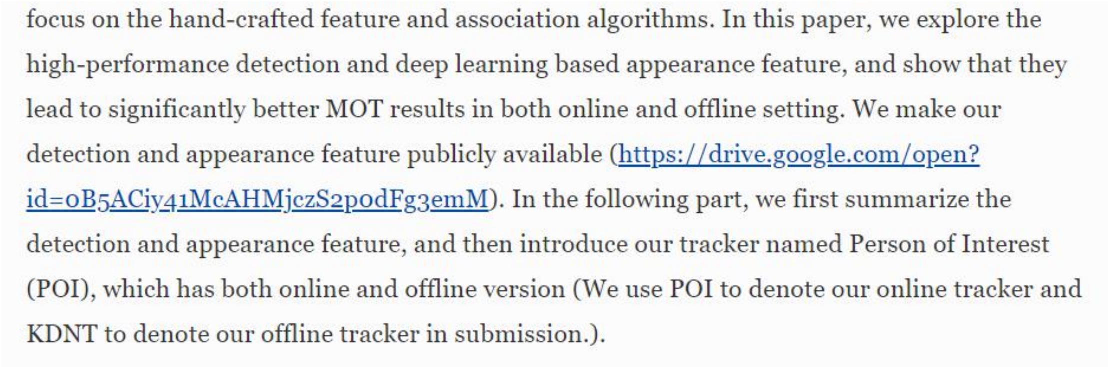
（e）代码（如果有写一下具体内容：实现、复现、配置运行？）

代码未开源

2. 文献 2

（a）文献名： F Yu, Li, W. , Li, Q. , Liu, Y. , Shi, X. , & Yan, J. . (2016). POI: Multiple Object Tracking with High Performance Detection and Appearance Feature. European Conference on Computer Vision. Springer, Cham.

（b）主要创新思想



focus on the hand-crafted feature and association algorithms. In this paper, we explore the high-performance detection and deep learning based appearance feature, and show that they lead to significantly better MOT results in both online and offline setting. We make our detection and appearance feature publicly available (https://drive.google.com/open?id=0B5ACiy41McAHMjczS2p0dFg3emM). In the following part, we first summarize the detection and appearance feature, and then introduce our tracker named Person of Interest (POI), which has both online and offline version (We use POI to denote our online tracker and KDNT to denote our offline tracker in submission.).

（c）主要原理剖析及说明

这篇文章的基本思路是在每帧上用检测器检测行人位置,在每帧之前利用行人检测框的表观特征(Appearance Feature)进行前后帧行人框的匹配,从而实现对行人的跟踪.所以这篇文章的算法算是 Tracking by Detection.这篇文章在行人检测器和表观特征提取两处均使用了基于深度学习的方法.并达到了较好的效果.作者给这个跟踪算法起了个叫 POI:Person of Interest.

POI 中使用检测器信息如下:

* 模型:Faster RCNN

* 数据库:使用了多个数据库.包括 ImageNet, ETHZ pedestrain dataset, Caltech pedestrain dataset 以及作者自己准备的数据集(接近 40w 个样本,但并没有公开).

* 策略:作者额外使用了 skip pooling [参考文献 1]和 multi-region[参考文献 2]这两个策略提高检测器的效果.

作者在 MOT16 train set 上对比了 Faster RCNN 和 DPM 的效果对比.其中添

加了 skip pooling 和 multi-region 两个策略的 Faster RCNN 获得了最好的综合效果，如下图所示.

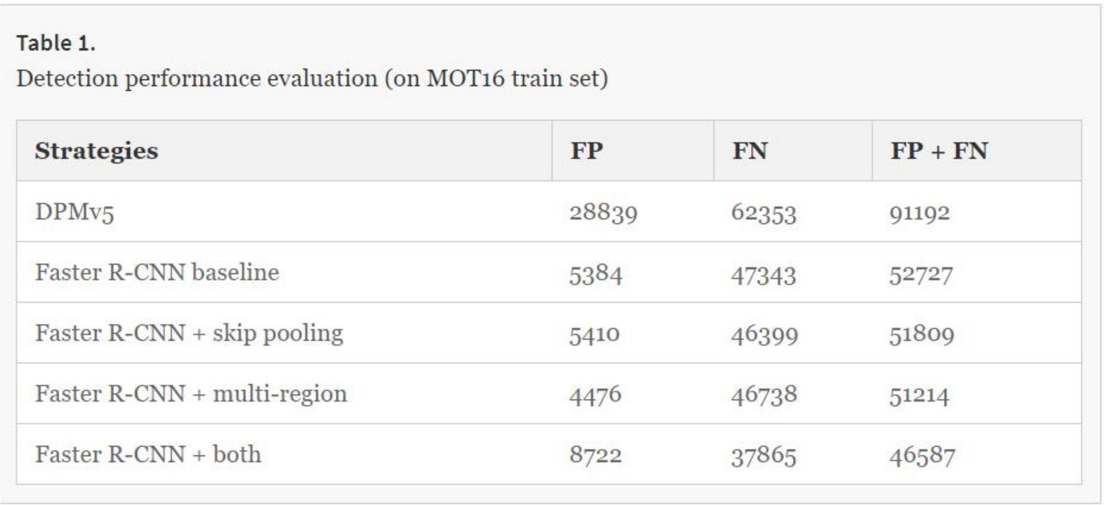


Table 1.
Detection performance evaluation (on MOT16 train set)

Strategies	FP	FN	FP + FN
DPMv5	28839	62353	91192
Faster R-CNN baseline	5384	47343	52727
Faster R-CNN + skip pooling	5410	46399	51809
Faster R-CNN + multi-region	4476	46738	51214
Faster R-CNN + both	8722	37865	46587

（d）主要实验结果（现有原文章中的）

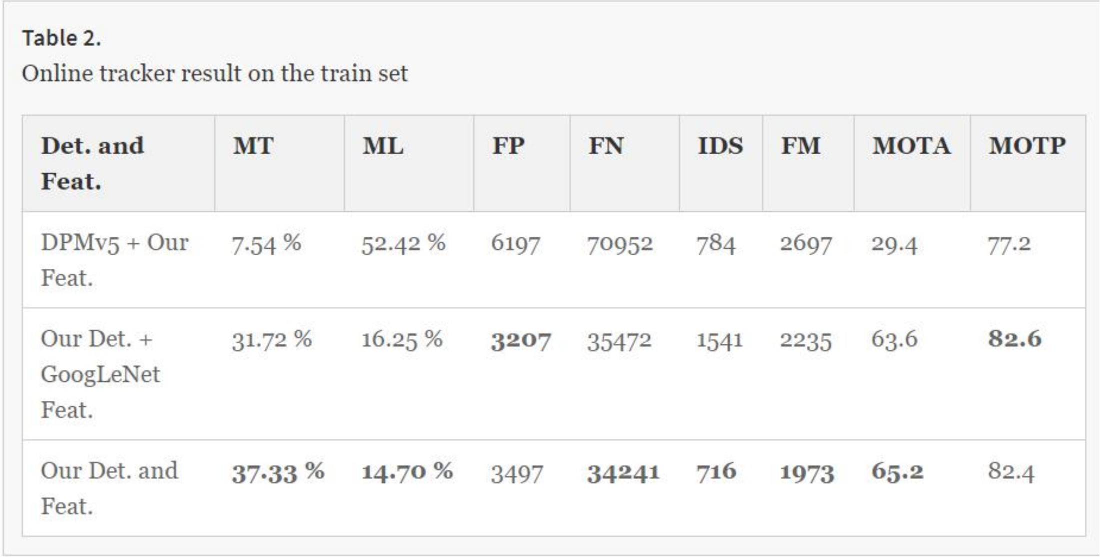


Table 2.
Online tracker result on the train set

Det. and Feat.	MT	ML	FP	FN	IDS	FM	MOTA	MOTP
DPMv5 + Our Feat.	7.54 %	52.42 %	6197	70952	784	2697	29.4	77.2
Our Det. + GoogLeNet Feat.	31.72 %	16.25 %	**3207**	35472	1541	2235	63.6	**82.6**
Our Det. and Feat.	**37.33 %**	**14.70 %**	3497	**34241**	**716**	**1973**	**65.2**	82.4

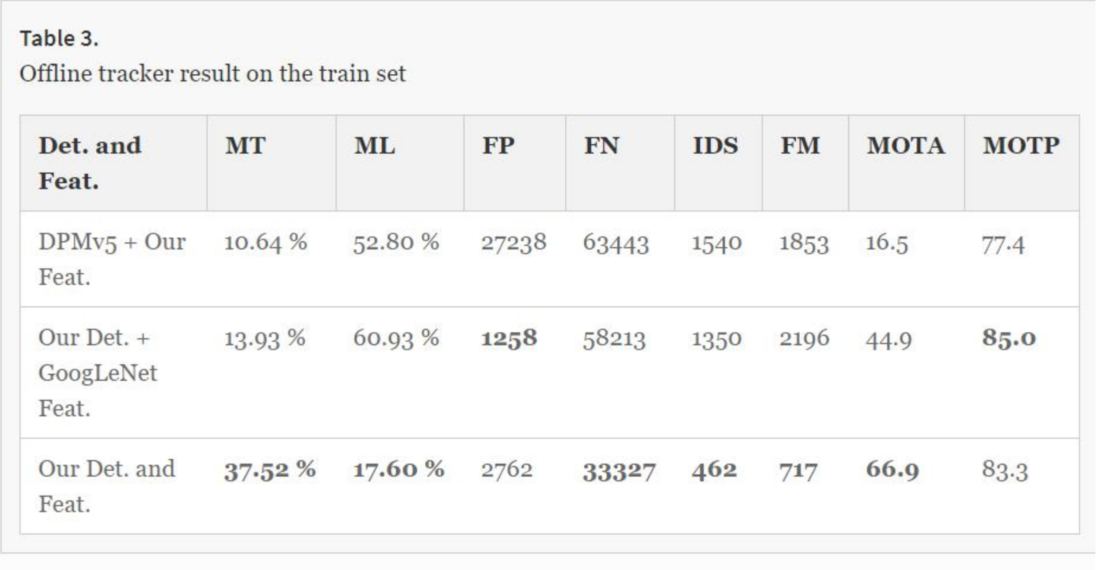


Table 3.
Offline tracker result on the train set

Det. and Feat.	MT	ML	FP	FN	IDS	FM	MOTA	MOTP
DPMv5 + Our Feat.	10.64 %	52.80 %	27238	63443	1540	1853	16.5	77.4
Our Det. + GoogLeNet Feat.	13.93 %	60.93 %	**1258**	58213	1350	2196	44.9	**85.0**
Our Det. and Feat.	**37.52 %**	**17.60 %**	2762	**33327**	**462**	**717**	**66.9**	83.3

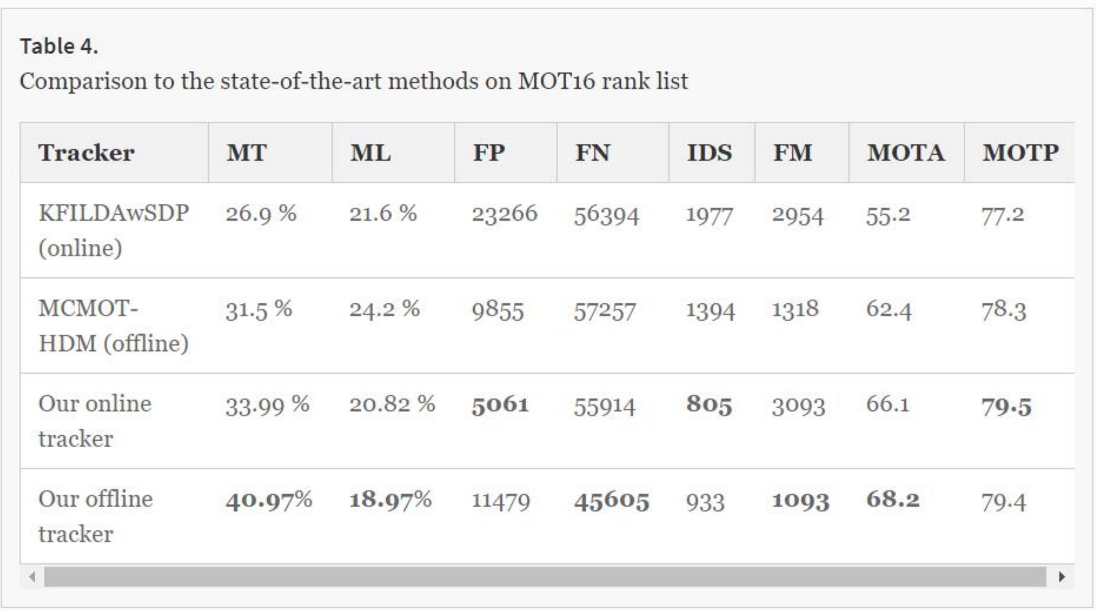


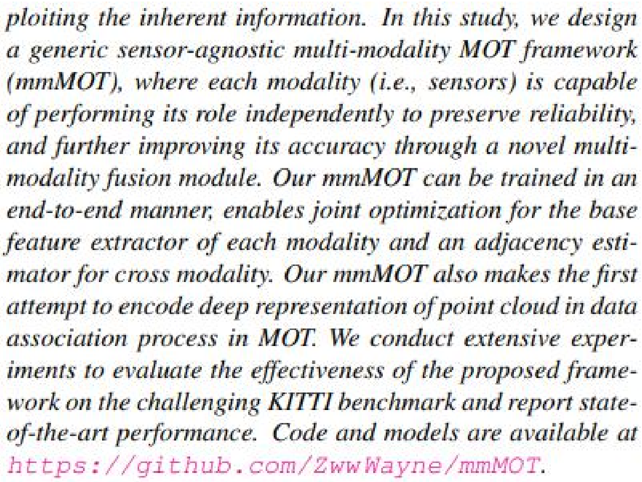
Table 4.
Comparison to the state-of-the-art methods on MOT16 rank list

Tracker	MT	ML	FP	FN	IDS	FM	MOTA	MOTP
KFILDAwSDP (online)	26.9 %	21.6 %	23266	56394	1977	2954	55.2	77.2
MCMOT-HDM (offline)	31.5 %	24.2 %	9855	57257	1394	1318	62.4	78.3
Our online tracker	33.99 %	20.82 %	**5061**	55914	**805**	3093	66.1	**79.5**
Our offline tracker	**40.97%**	**18.97%**	11479	**45605**	933	**1093**	**68.2**	79.4

3. 文献 3

（a）文献名： Zhang, W. , Zhou, H. , Sun, S. , Wang, Z. , Shi, J. , & Loy, C. C. . (2019). Robust Multi-Modality Multi-Object Tracking. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE.

（b）主要创新思想



ploiting the inherent information. In this study, we design a generic sensor-agnostic multi-modality MOT framework (mmMOT), where each modality (i.e., sensors) is capable of performing its role independently to preserve reliability, and further improving its accuracy through a novel multi-modality fusion module. Our mmMOT can be trained in an end-to-end manner, enables joint optimization for the base feature extractor of each modality and an adjacency estimator for cross modality. Our mmMOT also makes the first attempt to encode deep representation of point cloud in data association process in MOT. We conduct extensive experiments to evaluate the effectiveness of the proposed framework on the challenging KITTI benchmark and report state-of-the-art performance. Code and models are available at `https://github.com/ZwwWayne/mmMOT`.

1） 本文提出了一个多模态 MOT 框架，其中包含一个稳健的融合模块，利用多模态信息来提高可靠性和准确性。

2） 本文提出了一种新的端到端训练方法，使连接能够优化跨模态推理。

3） 本文首次尝试应用 pointcloud 的深层特征进行跟踪，并获得具有竞争力的结果。

（c）主要原理剖析及说明

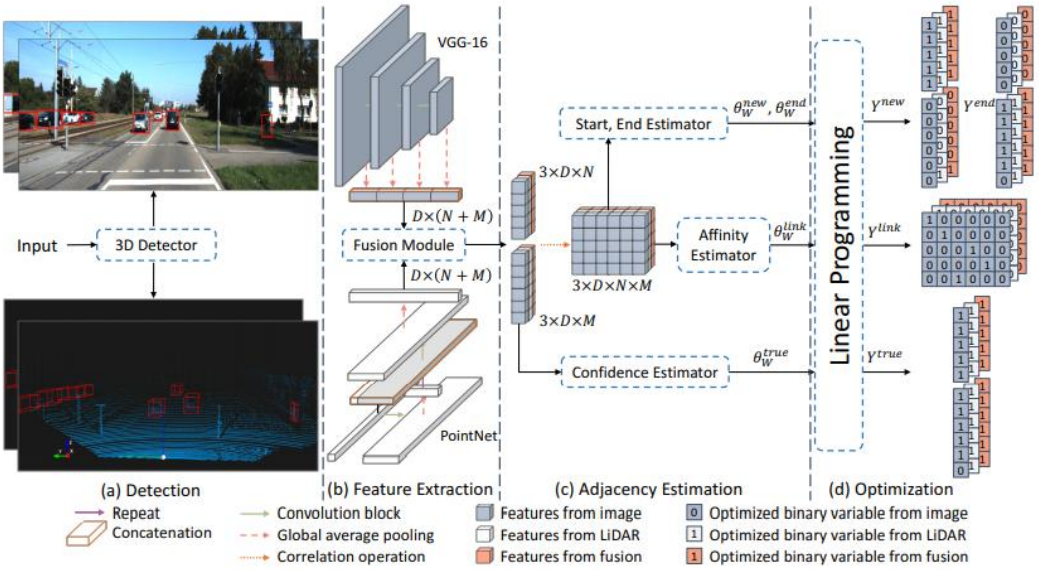


Figure 2. The pipeline of mmMOT. The feature extractors first extract features from image and LiDAR, and the robust fusion module fuses the multi-sensor features. Next, the correlation operator produces the correlation features for each detection pair, by which the adjacency estimator predicts the adjacency matrix. All the predicted scores are optimized to predict the binary variable Y.

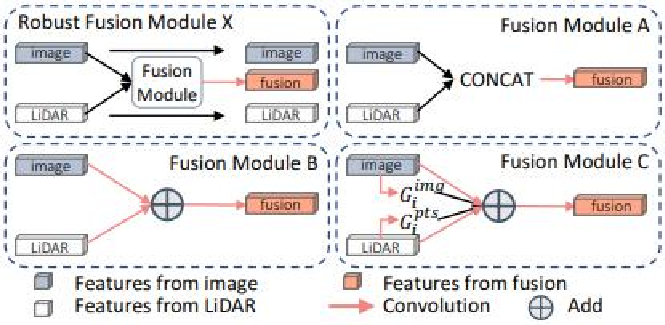
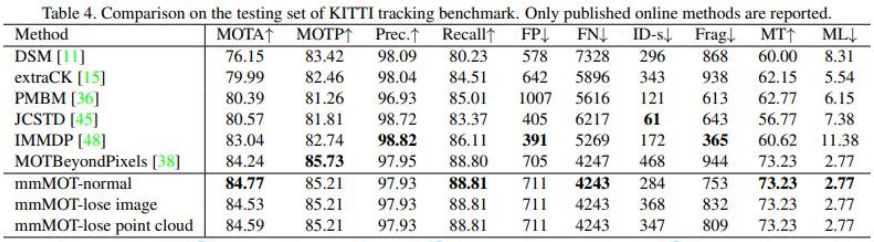


Figure 3. The robust fusion module and three multi-modality fusion modules. The robust fusion module can apply any one of the fusion modules A, B and C to produce the fused modality. Unlike the conventional fusion modules, the robust fusion module produces both the single modalities and the fused modality as an output. Fusion module A concatenates the multi-modality features, module B fuses them with a linear combination, module C introduces attention mechanism to weights the importance of sensor's feature adaptively.

（d）主要实验结果（现有原文章中的）

Table 4. Comparison on the testing set of KITTI tracking benchmark. Only published online methods are reported.



Method	MOTA↑	MOTP↑	Prec.↑	Recall↑	FP↓	FN↓	ID-s↓	Frag↓	MT↑	ML↓
DSM [11]	76.15	83.42	98.09	80.23	578	7328	296	868	60.00	8.31
extraCK [15]	79.99	82.46	98.04	84.51	642	5896	343	938	62.15	5.54
PMBM [36]	80.39	81.26	96.93	85.01	1007	5616	121	613	62.77	6.15
JCSTD [45]	80.57	81.81	98.72	83.37	405	6217	**61**	643	56.77	7.38
IMMDP [48]	83.04	82.74	**98.82**	86.11	**391**	5269	172	**365**	60.62	11.38
MOTBeyondPixels [38]	84.24	**85.73**	97.95	88.80	705	4247	468	944	73.23	2.77
mmMOT-normal	**84.77**	85.21	97.93	**88.81**	711	**4243**	284	753	**73.23**	**2.77**
mmMOT-lose image	84.53	85.21	97.93	88.81	711	4243	368	832	73.23	2.77
mmMOT-lose point cloud	84.59	85.21	97.93	88.81	711	4243	347	809	73.23	2.77

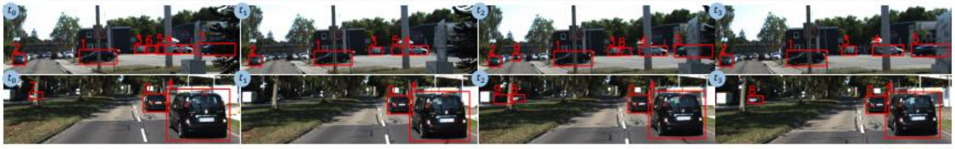


Figure 5. Failure case analysis.

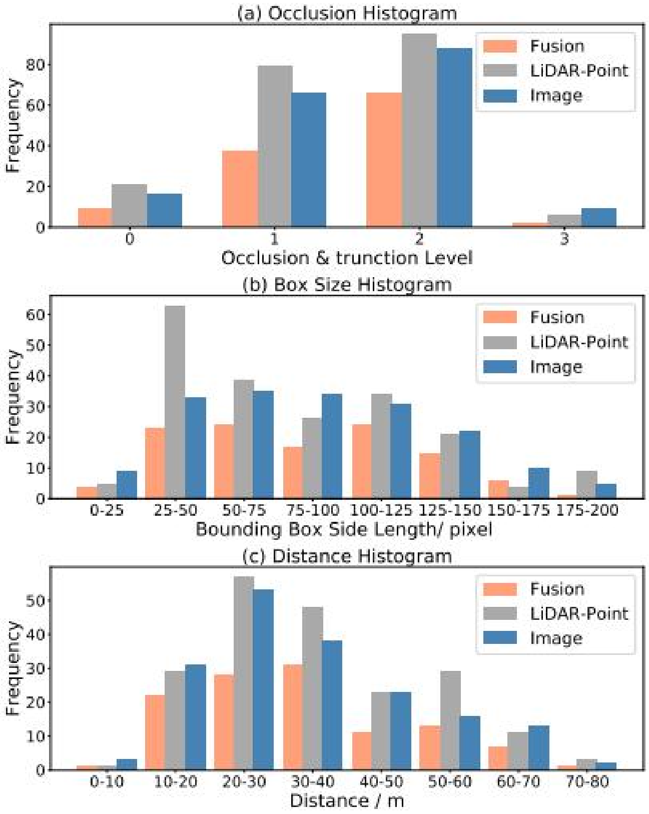


Figure A1. Failure case analysis. Occlusion level 0, 1, 2, 3 indicates the object is not, moderately, highly, extremely occluded and truncated in image.

三、本章学习小结

本章学习到了计算机视觉目标跟踪相关的专业知识。阅读了几篇论文后，对现在目标跟踪领域有一个大致的认识。还实现了部分的代码。