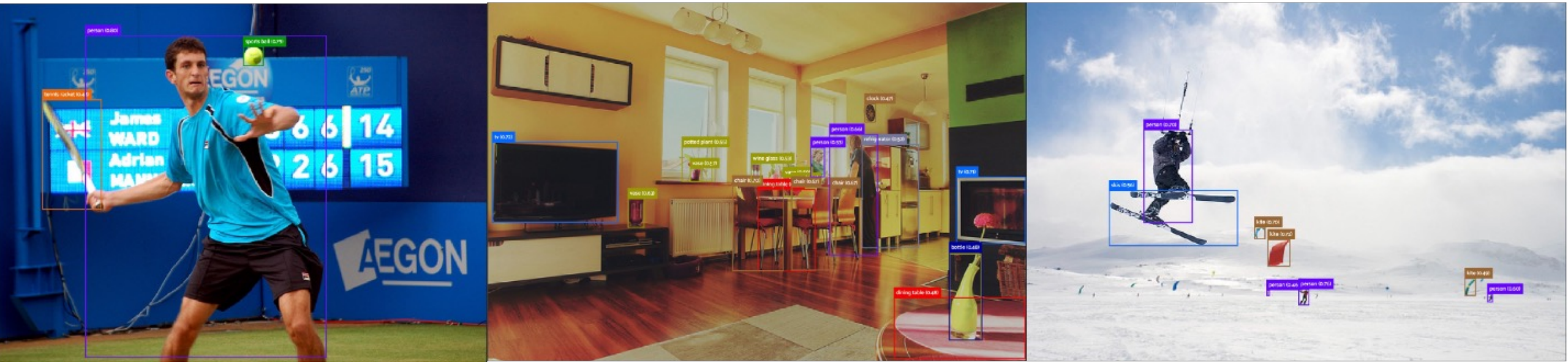
Alibaba Group 阿里巴巴集团

兼顾速度与精度的高效目标检测框架

DAMO-YOLO

from Alibaba Group



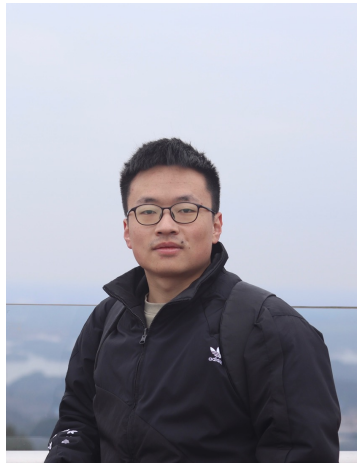
阿里巴巴达摩院智能计算实验室-TinyML团队

TinyML团队 DAMO-YOLO项目组

许贤哲

姜奕祺

陈威华



黄一伦

张袁

孙修宇

项目负责人

钉钉

微信

DAMO-YOLO : A Report on Real-Time Object Detection Design

Xianzhe Xu*, Yiqi Jiang*, Weihua Chen*, Yilun Huang*,Yuan Zhang*, Xiuyu Sun†
Alibaba Group

DAMO-YOLO

from Alibaba Group

- 目标检测简介
- 目标检测现状
- DAMO-YOLO技术价值
- DAMO-YOLO应用价值
- DAMO-YOLO原理简介

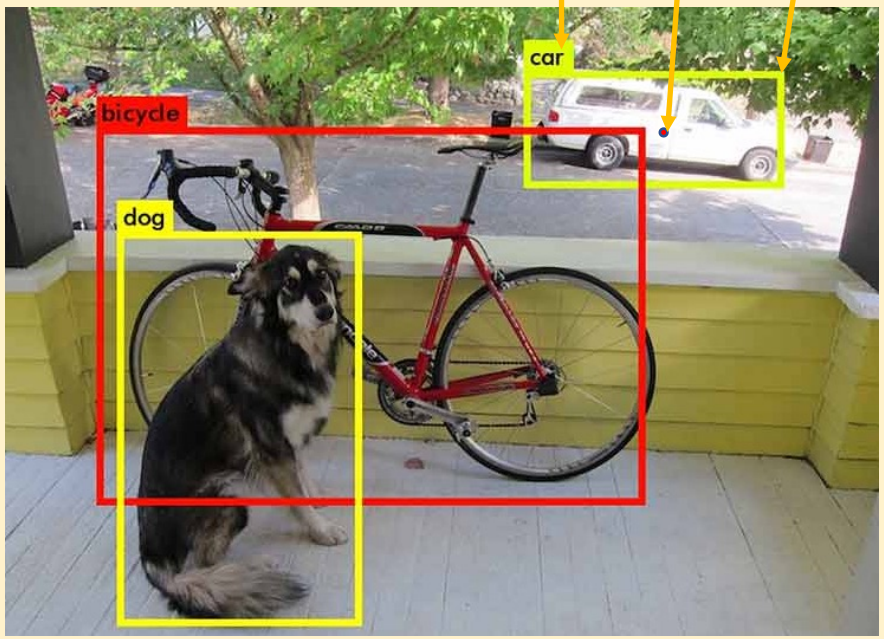
目标检测任务简介

定义：在图像/空间中定位出感兴趣物体的位置和大小

输入：图像/视频/点云

输出：物体类别 & 检测框坐标

类别



目标检测示意

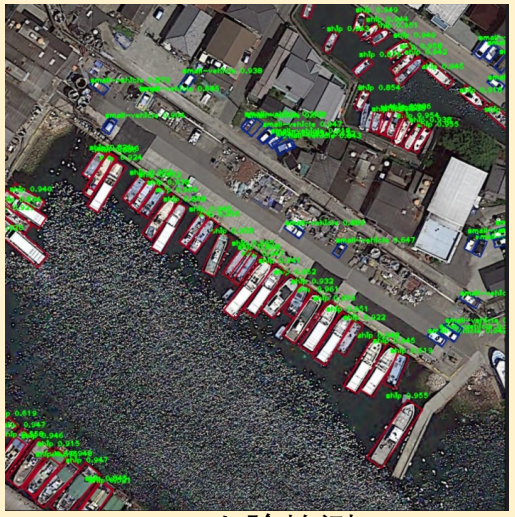
应用价值：

落地场景多样，CV应用的基础任务。

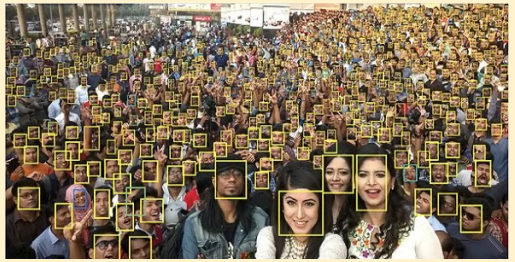
自动驾驶

侵入检测

码头管理



人脸检测



目标检测现状

现状：目标检测框架繁多，各有千秋。

baseline齐全，受学术欢迎

MMDetection

飞桨 PaddleDetection

3～5款模型(10~100GFlops)，
精简易用，工业部署友好

Detectron2

YOLOX
Exceeding YOLO series in 2021

YOLOv5

YOLOv6

灵活、
高效

DAMO-YOLO
from Alibaba Group

Github：https://github.com/tinyvision/DAMO-YOLO
Arxiv：https://arxiv.org/abs/2211.15444
ModelScope：
https://www.modelscope.cn/models/damo/cv_tinynas_object-detection_damoyolo/summary

现有框架应用痛点：

1.模型scale变化不够灵活，难以适应不同的算力场景
2.多尺度检测能力弱，难以适应不同的应用场景
3.精度-速度曲线不够理想，难以满足实时需求

DAMO-YOLO技术优势：

1. 整合自研NAS技术，可低成本自定义模型
2. EfficientRepGFPN+HeavyNeck，应用范围广
3. 全scale模型通用蒸馏技术，无损提升精度

DAMO-YOLO技术价值

DAMO-YOLO

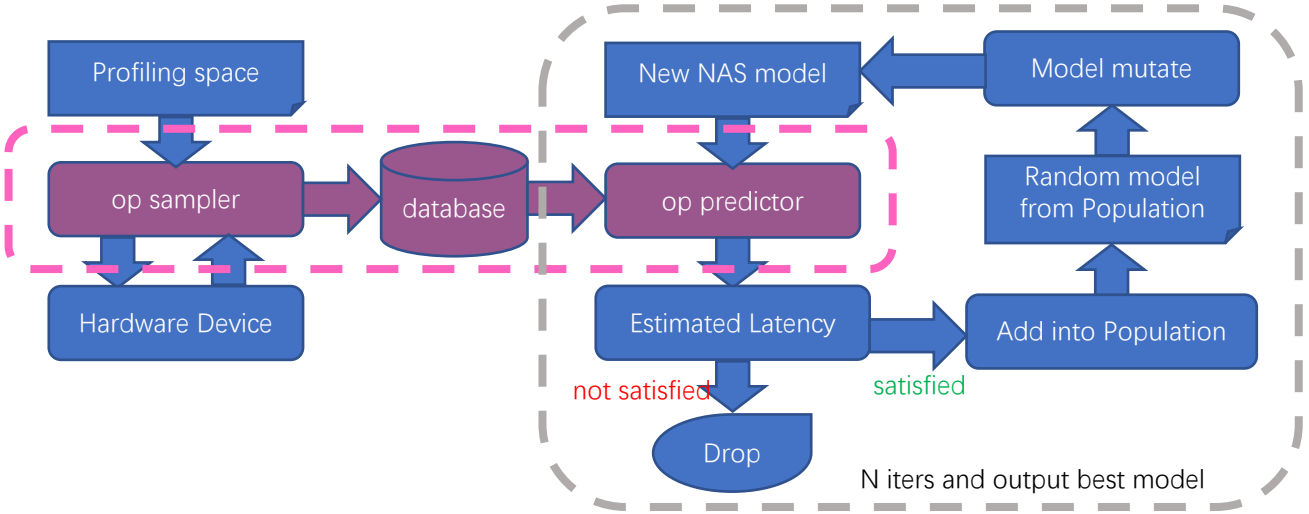
from Alibaba Group

DAMO-YOLO技术优势：

1. 整合自研NAS技术，可低成本自定义模型
2. EfficientRepGFPN+HeavyNeck，应用范围广
3. 全scale模型通用蒸馏技术，无损提升精度

低成本定制化模型能力：

- 基于自研MAE-NAS (ZeroShot)算法搜索最优模型
  - 无需训练，搜索成本低
  - 支持flops/latency作为预算搜索，提升芯片利用效率
- 提供了针对不同硬件的latency数据库构建方案
  - 支持T4/V100 GPU、IoT芯片等



基于lantecy进行模型搜索

DAMO-YOLO技术价值

DAMO-YOLO

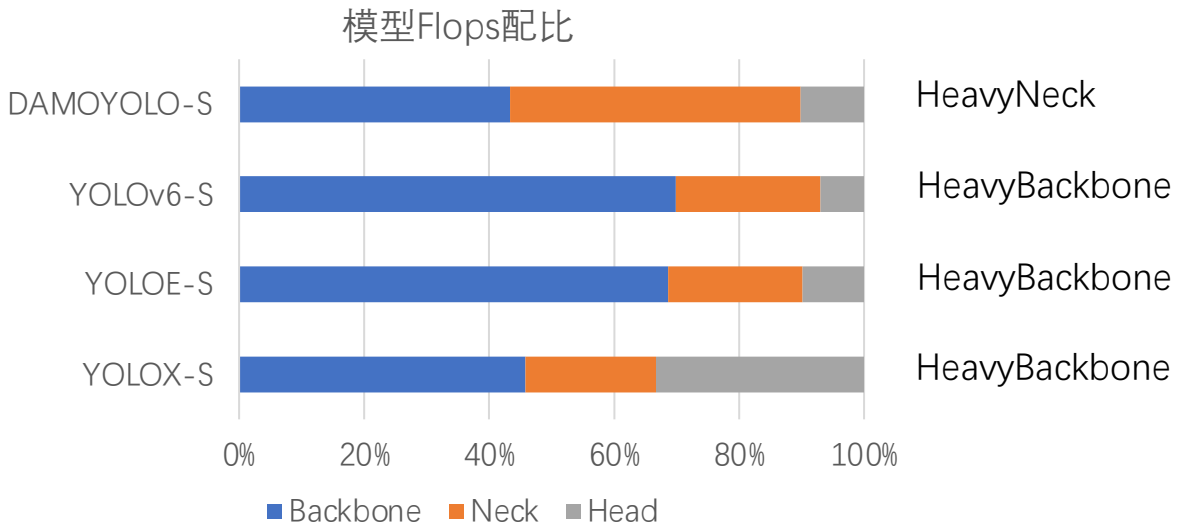
from Alibaba Group

DAMO-YOLO技术优势：

1. 整合自研NAS技术，可低成本自定义模型
2. EfficientRepGFPN+HeavyNeck，应用范围广
3. 全scale模型通用蒸馏技术，无损提升精度

EfficientRepGFPN+HeavyNeck，应用范围广

- Efficient RepGFPN，高效的多尺度特征融合
- HeavyNeck，重新定义模型配比
- 多尺度性能优异，模型应用范围广



DAMO-YOLO技术价值

DAMO-YOLO技术优势：

1. 整合自研NAS技术，可低成本自定义模型
2. EfficientRepGFPN+HeavyNeck，应用范围广
3. 全scale模型通用蒸馏技术，无损提升精度

全尺度模型通用蒸馏

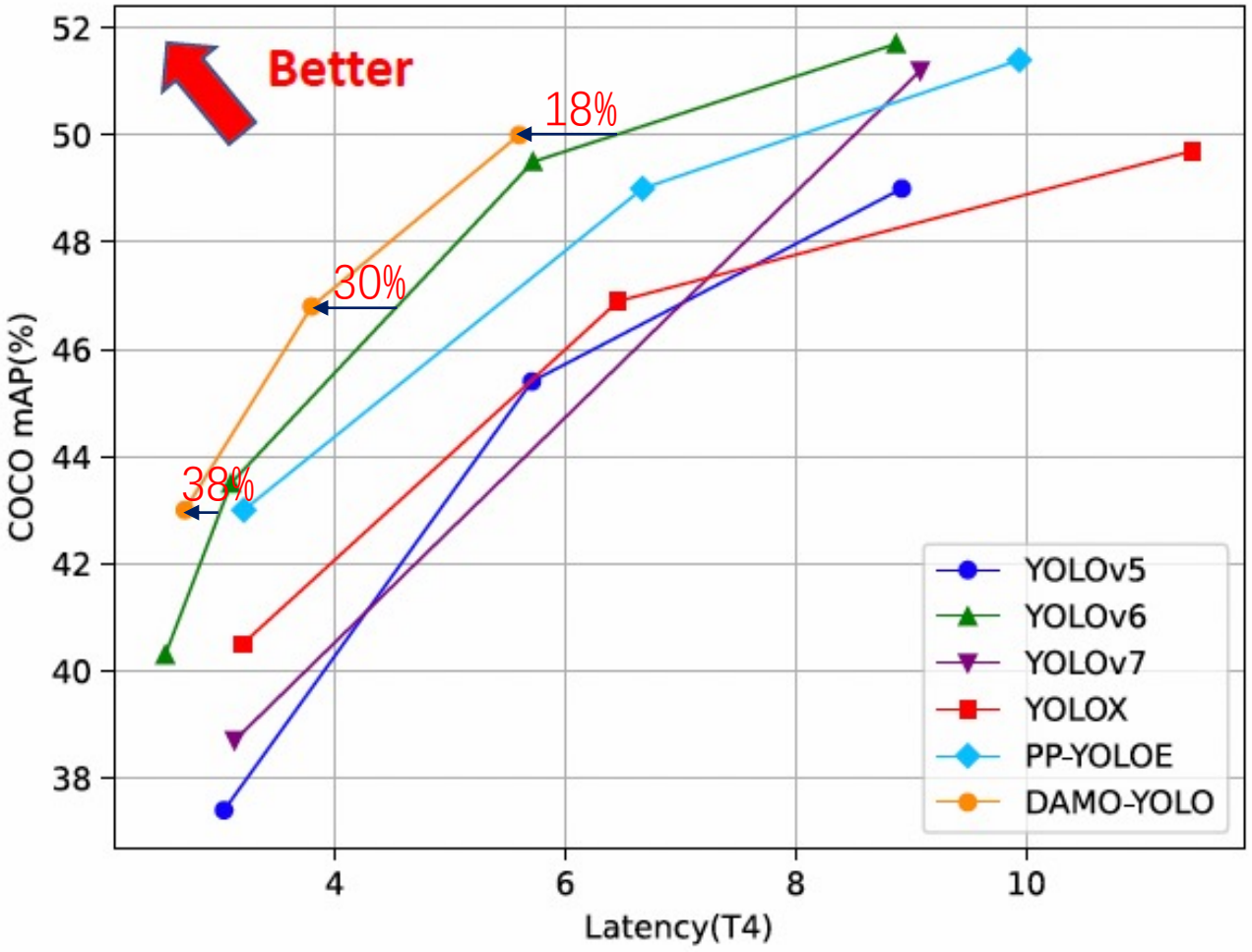
- 学术界和工业界对YOLO模型蒸馏探索较少
- 缺乏小模型蒸馏方案
  - FGD在YOLOX-M上蒸馏
  - YOLOv6在L、M上蒸馏
- DAMO-YOLO方案
  - 全尺度模型涨点明显。
  - 调参free，一键式脚本完成蒸馏。
  - 特征蒸馏+无偏BN+AlignModule，异构鲁棒。

全尺度模型蒸馏在T/S/M的效果

Method	Size	Latency(ms)	GFLOPs	Params(M)	AP
DAMO-YOLO-T	640	2.78	18.1	8.5	41.8
DAMO-YOLO-T*	640	2.78	18.1	8.5	43.0
DAMO-YOLO-S	640	3.83	37.8	16.3	45.6
DAMO-YOLO-S*	640	3.83	37.8	16.3	46.8
DAMO-YOLO-M	640	5.62	61.8	28.2	48.7
DAMO-YOLO-M*	640	5.62	61.8	28.2	50.0

DAMO-YOLO应用价值

检测框架性能对比



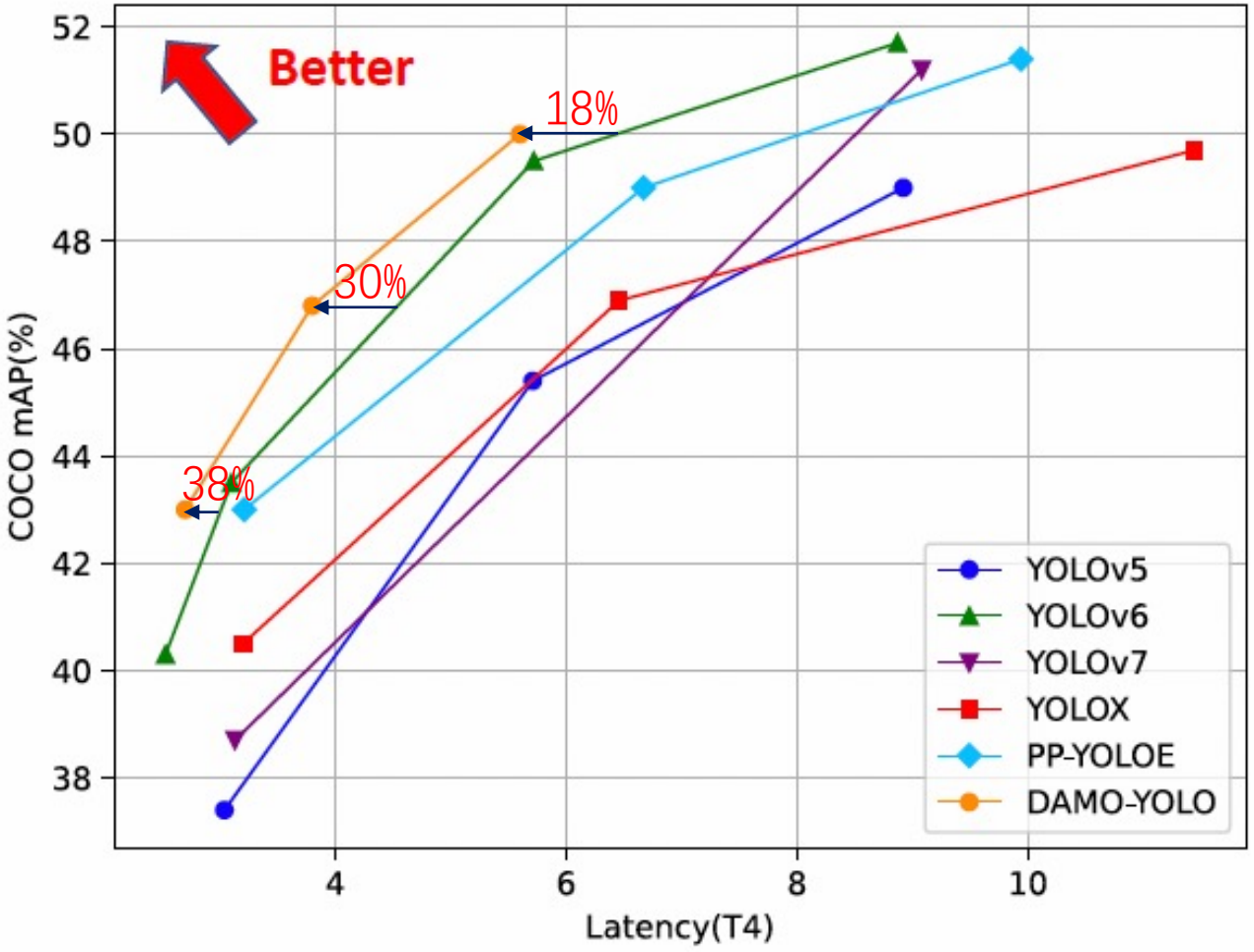
1. 模型提速20% ~ 40%
2. 计算量(Flops)减少15% ~ 50%
3. 参数量减少6%~50%
4. 全尺度涨点明显，适用范围广

Method	Size	Latency(ms)	GFLOPs	Params(M)	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}
YOLOX-T	416	1.78	6.5	5.1	32.8	-	-	-	-	-
YOLOX-S	640	3.20	26.8	9.0	40.5	-	-	-	-	-
YOLOX-M	640	6.46	73.8	25.3	46.9	-	-	-	-	-
YOLOX-L	640	11.44	155.6	54.2	49.7	-	-	-	-	-
YOLOv5-N	640	2.23	4.5	1.9	28.0	45.7	-	-	-	-
YOLOv5-S	640	3.04	16.5	7.2	37.4	56.8	-	-	-	-
YOLOv5-M	640	5.71	49.0	21.2	45.4	64.1	-	-	-	-
YOLOv5-L	640	8.92	109.1	46.5	49.0	67.3	-	-	-	-
YOLOv6-T	640	2.53	36.7	15.0	40.3	56.6	-	-	-	-
YOLOv6-S	640	3.10	44.2	17.0	43.5	60.4	-	-	-	-
YOLOv6-M*	640	5.72	82.2	34.3	49.5	66.8	-	-	-	-
YOLOv6-L*	640	9.87	144.0	58.5	52.5	70.0	-	-	-	-
YOLOv7-T-silu	640	3.13	13.7	6.2	38.7	56.7	41.7	18.8	42.4	51.9
YOLOv7	640	9.08	104.7	36.9	51.2	69.7	55.9	31.8	55.5	65.0
YOLOE-S	640	3.21	17.4	7.9	43.0	60.5	46.6	23.2	46.4	56.9
YOLOE-M	640	6.67	49.9	23.4	49.0	66.5	53.0	28.6	52.9	63.8
YOLOE-L	640	9.94	110.1	52.2	51.4	68.9	55.6	31.4	55.3	66.1
DAMO-YOLO-T	640	2.78	18.1	8.5	41.8	58.0	45.2	23.0	46.1	58.5
DAMO-YOLO-T*	640	2.78	18.1	8.5	43.0	59.4	46.6	23.3	47.4	61.0
DAMO-YOLO-S	640	3.83	37.8	16.3	45.6	61.9	49.5	25.9	50.6	62.5
DAMO-YOLO-S*	640	3.83	37.8	16.3	46.8	63.5	51.1	26.9	51.7	64.9
DAMO-YOLO-M	640	5.62	61.8	28.2	48.7	65.5	53.0	29.7	53.1	66.1
DAMO-YOLO-M*	640	5.62	61.8	28.2	50.0	66.8	54.6	30.4	54.8	67.6

SOTA模型对比

DAMO-YOLO应用价值

检测框架性能对比



DAMO-YOLO
from Alibaba Group

降本、增效

1.模型速度快，Flops低，适用范围广。
2.可针对算力自定义模型，提高芯片利用效率。
3.DAMO-YOLO-S已上线ModelScope！方便易用

```
from modelscope.pipelines import pipeline
from modelscope.utils.constant import Tasks
object_detect = pipeline(Tasks.image_object_detection,model='damo/cv_tinynas_object-detection_damoyolo')
img_path ='https://modelscope.oss-cn-beijing.aliyuncs.com/test/images/image_detection.jpg'
result = object_detect(img_path)
```

Github：https://github.com/tinyvision/DAMO-YOLO
Arxiv：https://arxiv.org/abs/2211.15444
ModelScope：https://www.modelscope.cn/models/damo/cv_tinynas_object-detection_damoyolo/summary

Alibaba Group 阿里巴巴集团

DAMO-YOLO

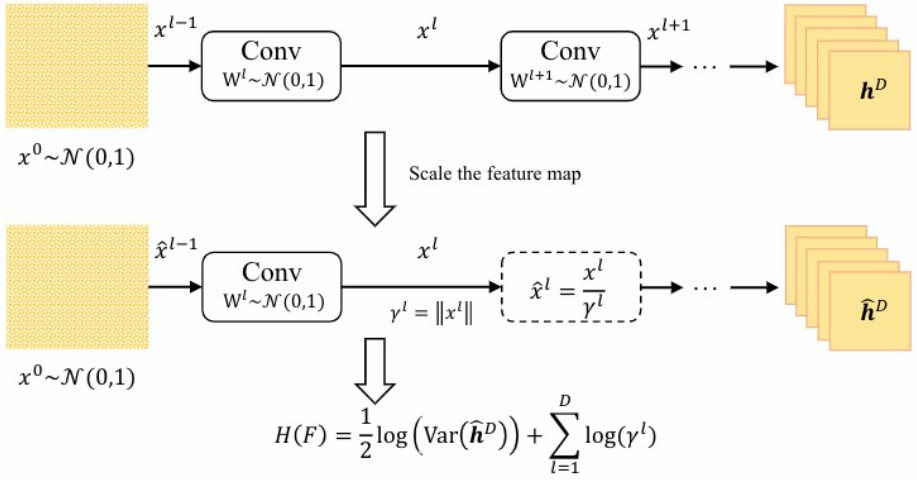
from Alibaba Group

- DAMO-YOLO原理简介
  - **低成本模型定制——MAE-NAS**
  - **高效多尺度融合——Efficient RepGFPN**
  - **全尺度模型蒸馏**

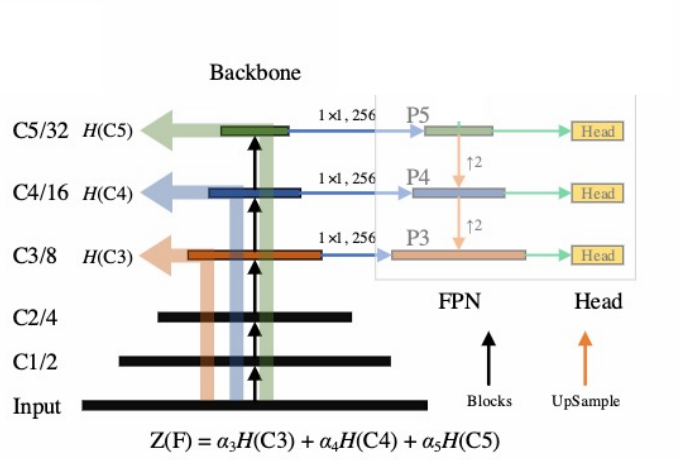
DAMO-YOLO基本原理

低成本模型定制——MAE-NAS （ICML2022）

- 基本思想：将深度网络视为连续状态空间的信息系统，并最大化该信息系统的熵
- 网络建模：
  - 将网络F的拓扑结构抽象为图$G=(V,E)$，顶点V表示特征，边E表示各种算子
  - $h(v)$和$h(e)$分别表示顶点和边中的值，$S=\{h(v),h(e):\forall v\in V,\forall e\in E\}$定义了网络$F$的连续状态空间
  - 熵$H(S)$衡量整个信息系统F包含的总信息量
    - 我们只关注网络的表达能力（顶点中的信息量），即$H(S_v)$
- 理论原理：
  - 根据**高斯分布微分熵**以及**高斯熵上界定理**，可计算特征图的方差来近似网络的熵$H(S_v)$，且只要其服从高斯分布就可估算出它的上界
  - 实际做法：
    - 用标准高斯分布$N(0,1)$初始化backbone参数并输入随机采样的标准高斯噪声
    - 网络F的单尺度（高斯上界）熵可表示为：$H(F)=\frac{1}{2}\log\left(Var(\hat{\boldsymbol{h}}^D)\right)+\sum_{l=1}^{D}\log(\gamma^l)$
    - 网络F的多尺度熵：$Z(F):=\alpha_1H(C1)+\alpha_2H(C2)+\cdots+\alpha_5H(C5)$ [0,0,1,1,6]



(a) Single-scale entropy score with rescaling

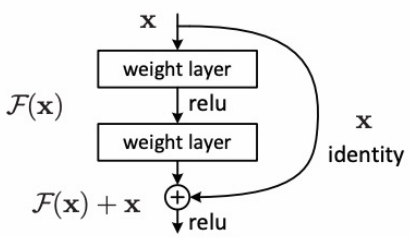


(b) Multi-scale entropy score for detection

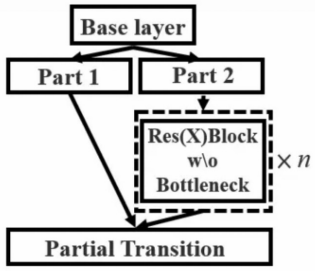
DAMO-YOLO基本原理

低成本模型定制—— MAE-NAS （ICML2022）

- NAS框架：Evolution算法
  - 以网络的多尺度熵作为性能proxy
  - **支持多种推理budget**：FLOPs，Params，Latency，Layers等
  - **支持细粒度mutate**：kernel size，width，depth，block类型等
  - **高可拓展性**：轻松根据官方教程添加用户自定义的结构block类型
  - **Zero-shot & GPU-free**：无需任何数据，无需GPU，只在CPU上最短几十分钟即可搜索出结果
- MAE-NAS Backbone for DAMO-YOLO
  - 针对实际应用时推理速度优先的目标，以不同latency限制搜索T/S/M模型。
  - 对搜索出的基础结构进行包装，小模型使用ResStyle，大模型使用CSPStyle。



ResStyle



CSPStyle

	Backbone	AP	Latency(ms)
DAMO-YOLO-S	CSP-Darknet	44.9	3.92
DAMO-YOLO-S	MAE-ResNet	45.6	3.83
DAMO-YOLO-S	MAE-CSP	45.3	3.79
DAMO-YOLO-M	MAE-ResNet	48.0	5.64
DAMO-YOLO-M	MAE-CSP	48.7	5.60

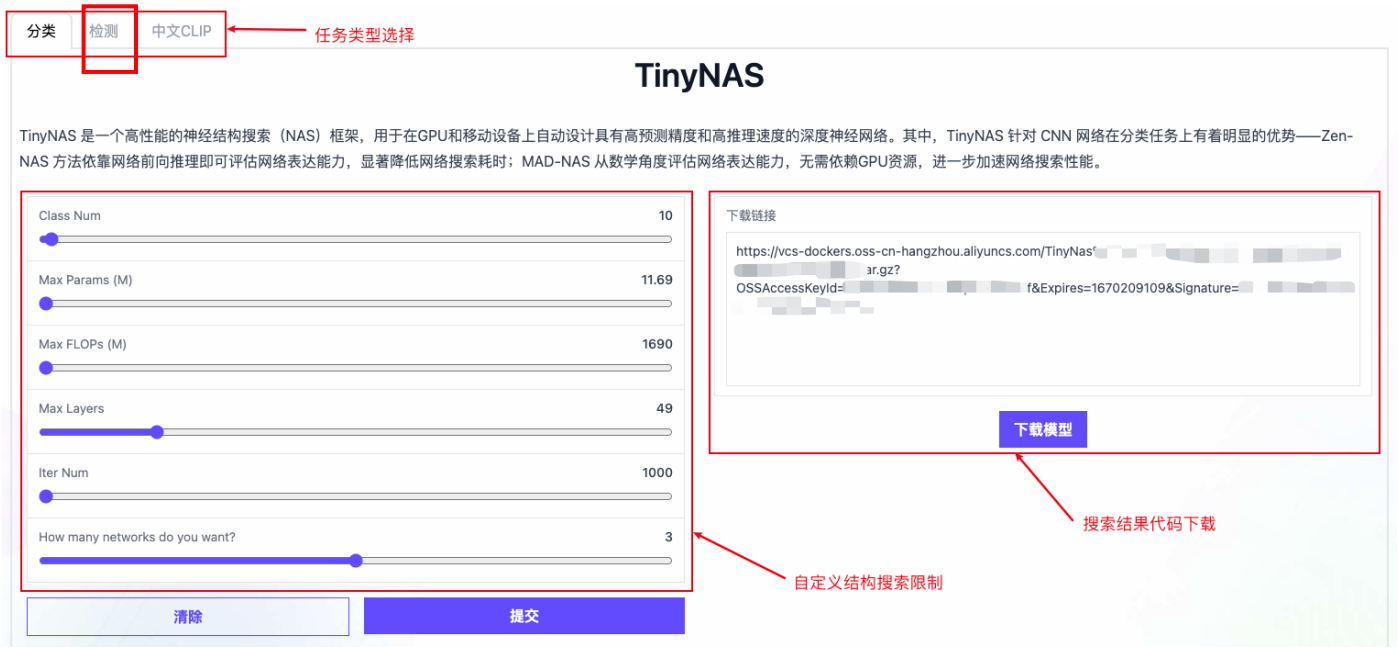
论文：ICML2022, *MAE-DET: Revisiting Maximum Entropy Principle in Zero-Shot NAS for Efficient Object Detection*

NAS for DAMO-YOLO教程： https://github.com/alibaba/lightweight-neural-architecture-search/blob/main/scripts/damo-yolo/Tutorial_NAS_for_DAMO-YOLO_cn.md

DAMO-YOLO基本原理

TinyNAS搜索工具现已上线ModelScope!

- 基于zero/one-shot方法，最短几分钟即可完成搜索流程
- 支持多场景任务：分类，检测，中文CLIP
- 支持结构搜索限制自定义：参数量，计算量，网络层数等
- 轻松集成搜索结果



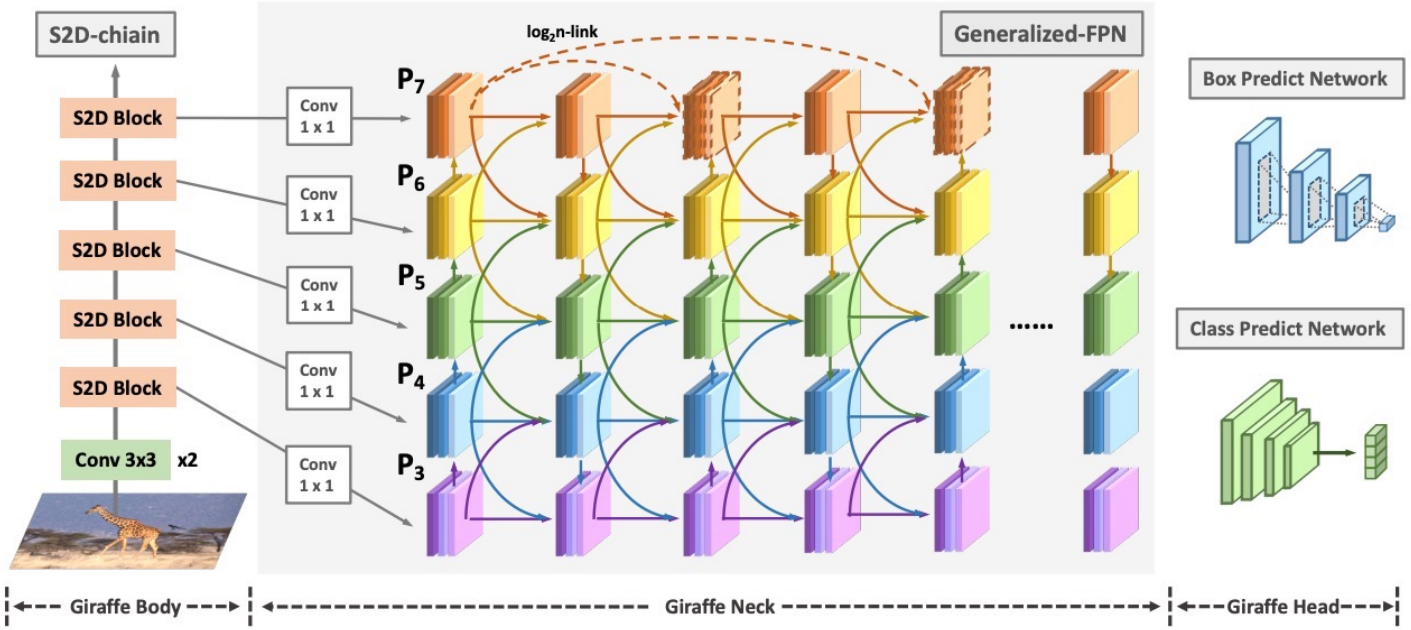
Github： https://github.com/alibaba/lightweight-neural-architecture-search

ModelScope: https://modelscope.cn/studios/damo/TinyNAS/summary

DAMO-YOLO基本原理

提升多尺度检测能力——GFPN（ICLR2022）

- 多尺度检测能力依赖于不同尺度特征的融合
- GFPN以相同优先级处理高层语义信息和低层空间信息，有益于多尺度特征融合互补
- 特征复用和更多连接提升了精度，但是并行化效率低，Flops高效Latency低效。



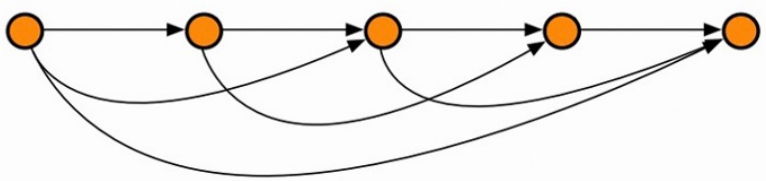
GFPN结构示意图

- 论文：*GiraffeDET: A Heavy-Neck Paradigm for Object Detection*, arXiv
- 代码：https://github.com/damo-cv/GiraffeDet

Skip Layer

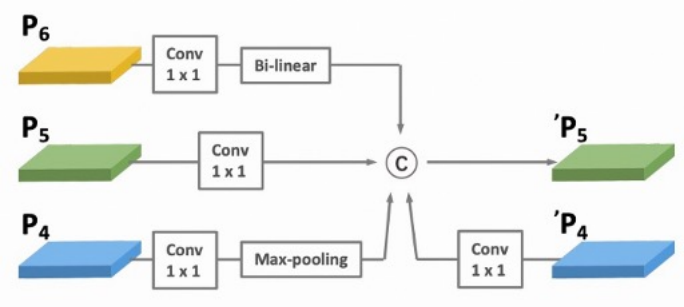
- log2n-link的连接增强特征复用，减少冗余

$$P_k^l = Conv(Concat(P_k^{l-2^n}, \ldots, P_k^{l-2^1}, P_k^{l-2^0})),$$



Queen Fusion

- 接收更多节点输入，增强特征表达能力
  - 斜上、斜下、串行连接



DAMO-YOLO基本原理

GFPN（ICLR2022） → **Efficient RepGFPN**

存在问题

- 不同尺度共享统一通道数
- Queen-Fusion带来了低效连接

- 节点堆叠运算效率低

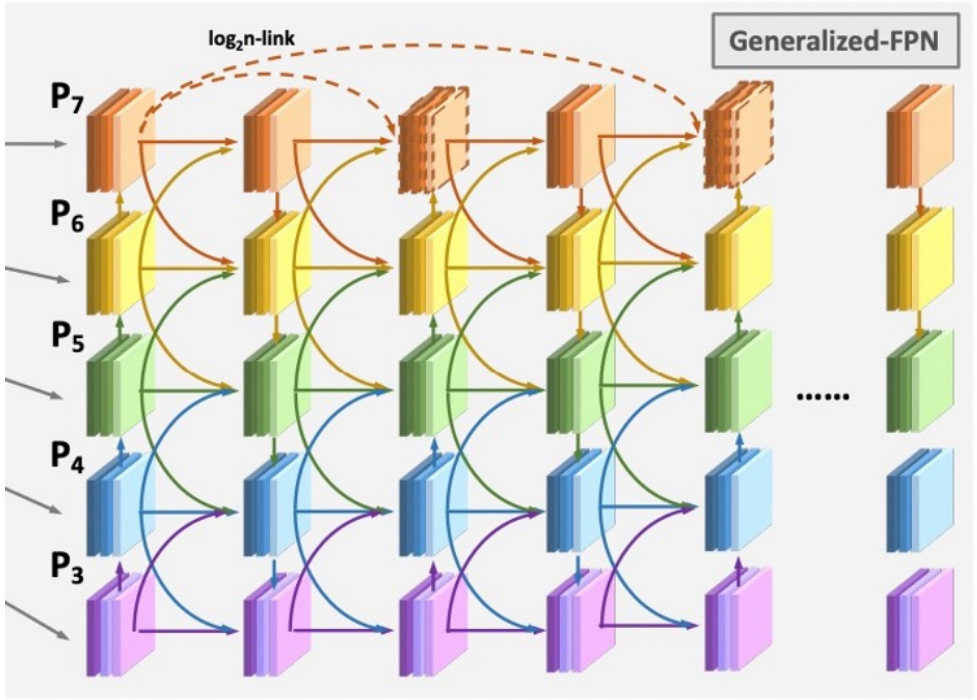
优化策略

拓扑结构优化 →

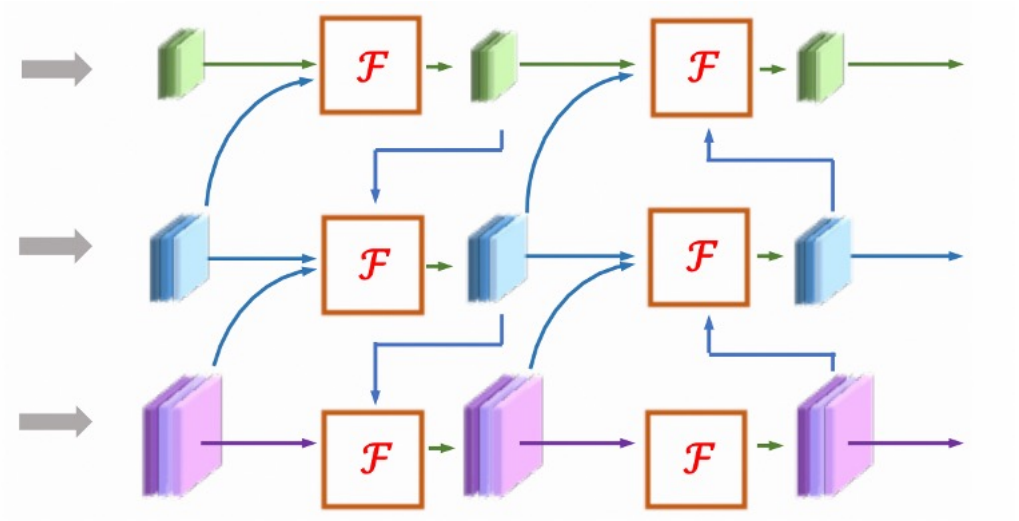
- 不同尺度特征使用不同的通道数
- 移除Queen-Fusion中的低效上采样算子

融合方式优化 →

- 固定节点数目，优化融合方式



GFPN结构示意图

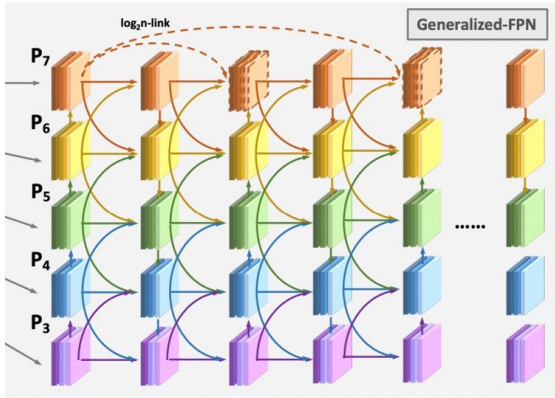


Efficient RepGFPN结构示意图

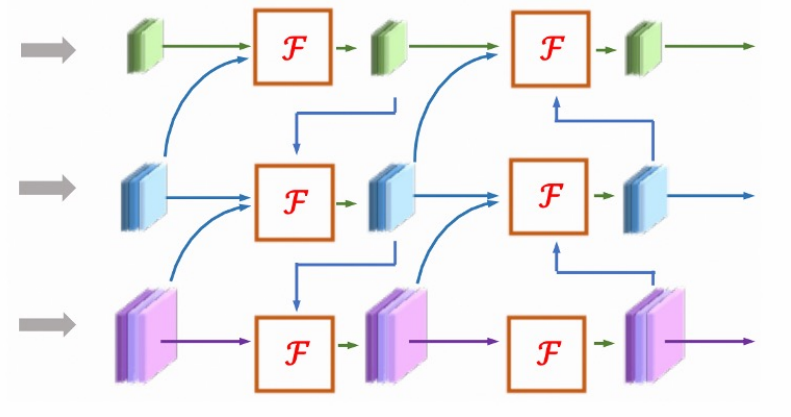
DAMO-YOLO基本原理

GFPN（ICLR2022）➡ **Efficient RepGFPN**

- **拓扑结构优化**
  - 不同尺度共享统一通道数➡不同尺度特征使用不同的通道数
  - Queen-Fusion带来了冗余连接➡移除Queen-Fusion中的额外的上采样算子



GFPN结构示意图



Efficient RepGFPN结构示意图

Depth	Width	Latency	FLOPs	AP
2	(192, 192, 192)	3.53	34.9	44.2
2	(128, 256, 512)	3.72	36.1	45.1
3	(160, 160, 160)	3.91	38.2	44.9
3	**(96, 192, 384)**	**3.83**	**37.8**	**45.6**
4	(64, 128, 256)	3.85	37.2	45.3

Efficient RepGFPN 深度与宽度分析

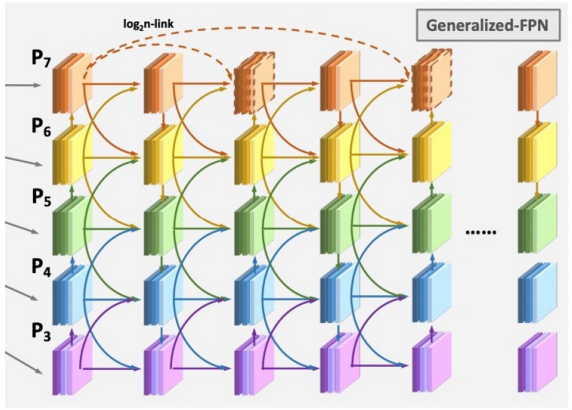
↘	↗	Latency	FLOPs	AP
		3.62	33.3	44.2
✓		4.19	37.7	44.5
	✓	**3.83**	**37.8**	**45.6**
✓	✓	4.58	42.8	45.9

Queen-Fusion连接效率分析

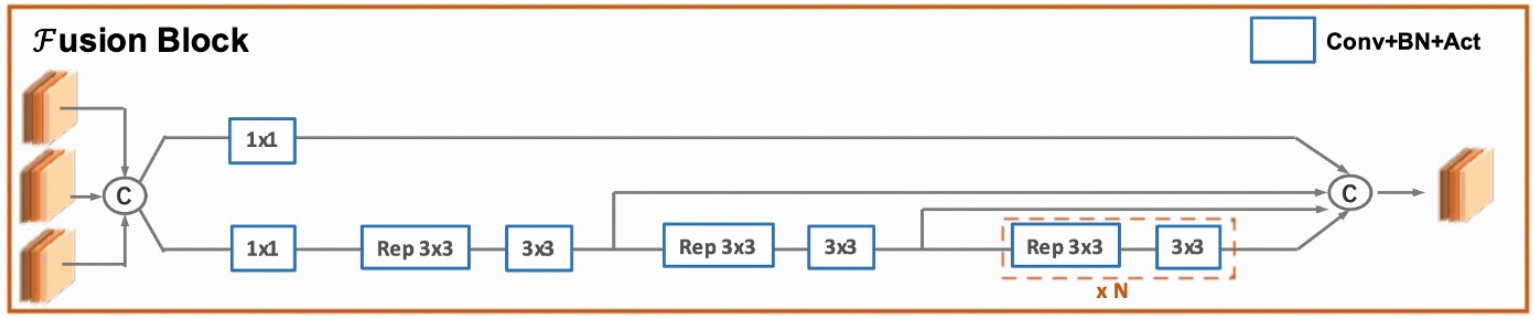
DAMO-YOLO基本原理

GFPN（ICLR2022）➡ Efficient RepGFPN

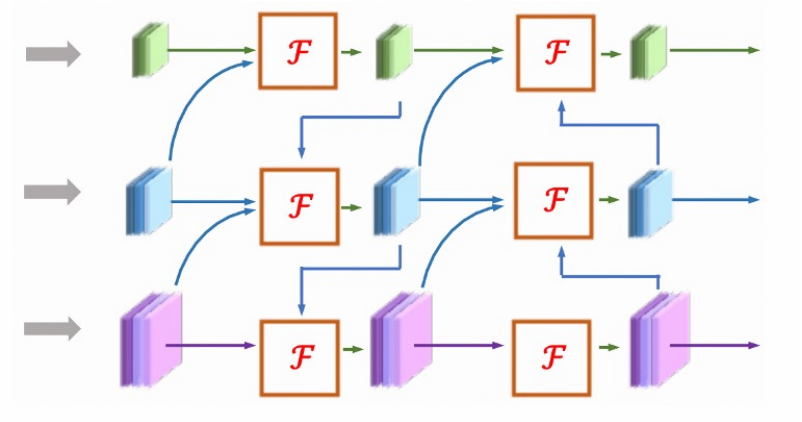
- **融合方式优化**
  - 节点堆叠运算效率低➡固定节点数目，FusionBlock融合
  - CSP连接、Rep重参数化机制、多层聚合连接



GFPN结构示意图



Fusion Block结构示意图



Efficient RepGFPN结构示意图

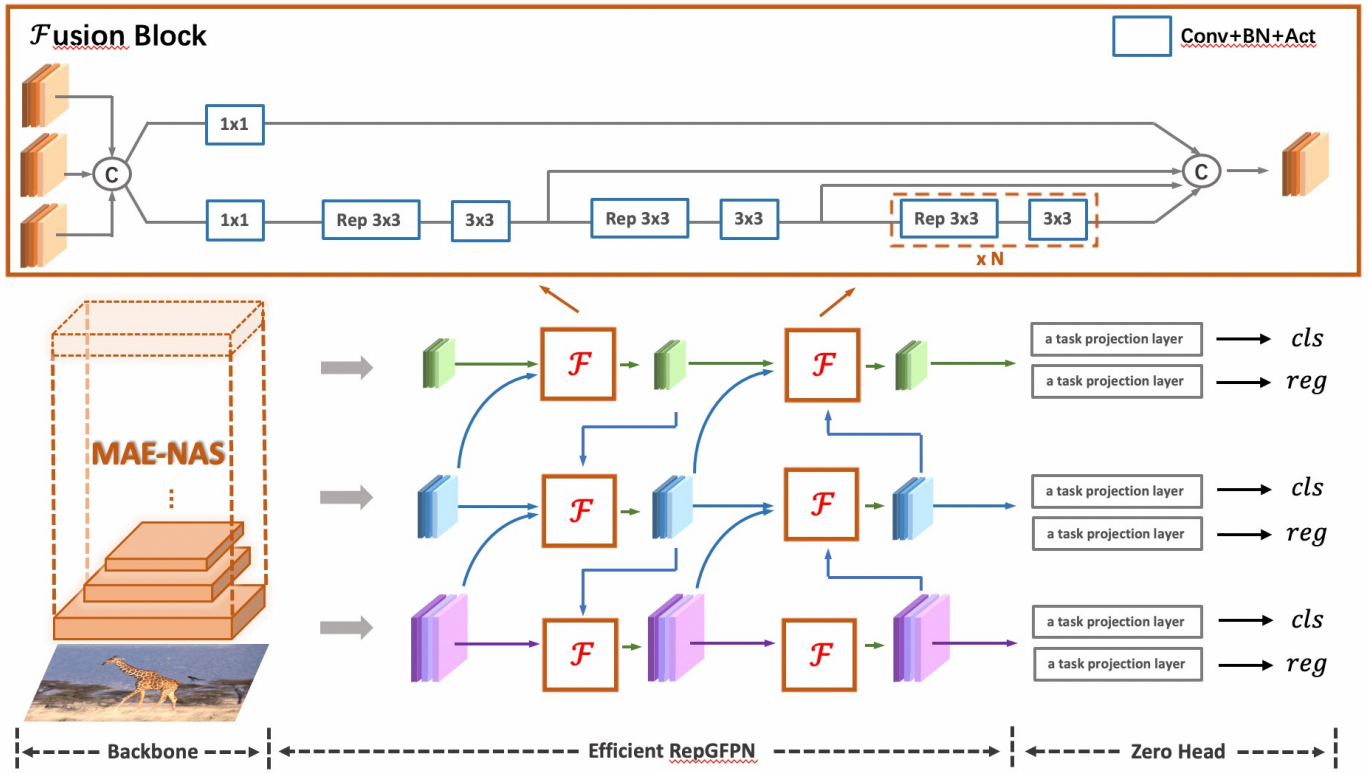
Merge−Style	Latency	FLOPs	AP
Conv	3.64	44.3	40.2
CSP	3.72	36.7	44.4
CSP + Reparam	3.72	36.7	45.0
CSP + Reparam + ELAN	**3.83**	**37.8**	**45.6**

DAMO-YOLO基本原理

HeavyNeck & ZeroHead

- Head只保留用于分类和回归任务的一层线性投影层
- 更多的计算量用于进行Efficient RepGFPN融合模块的堆叠

Neck(width/depth)	Head(width/depth)	Latency(ms)	AP
(1.0/1.0)	**(1.0/0.0)**	**3.83**	**45.6**
(1.0/0.50)	(1.0/1.0)	3.79	44.9
(1.0/0.33)	(1.0/2.0)	3.85	43.7
(1.0/0.0)	(1.0/3.0)	3.87	41.2



DAMO-YOLO模型结构示意图

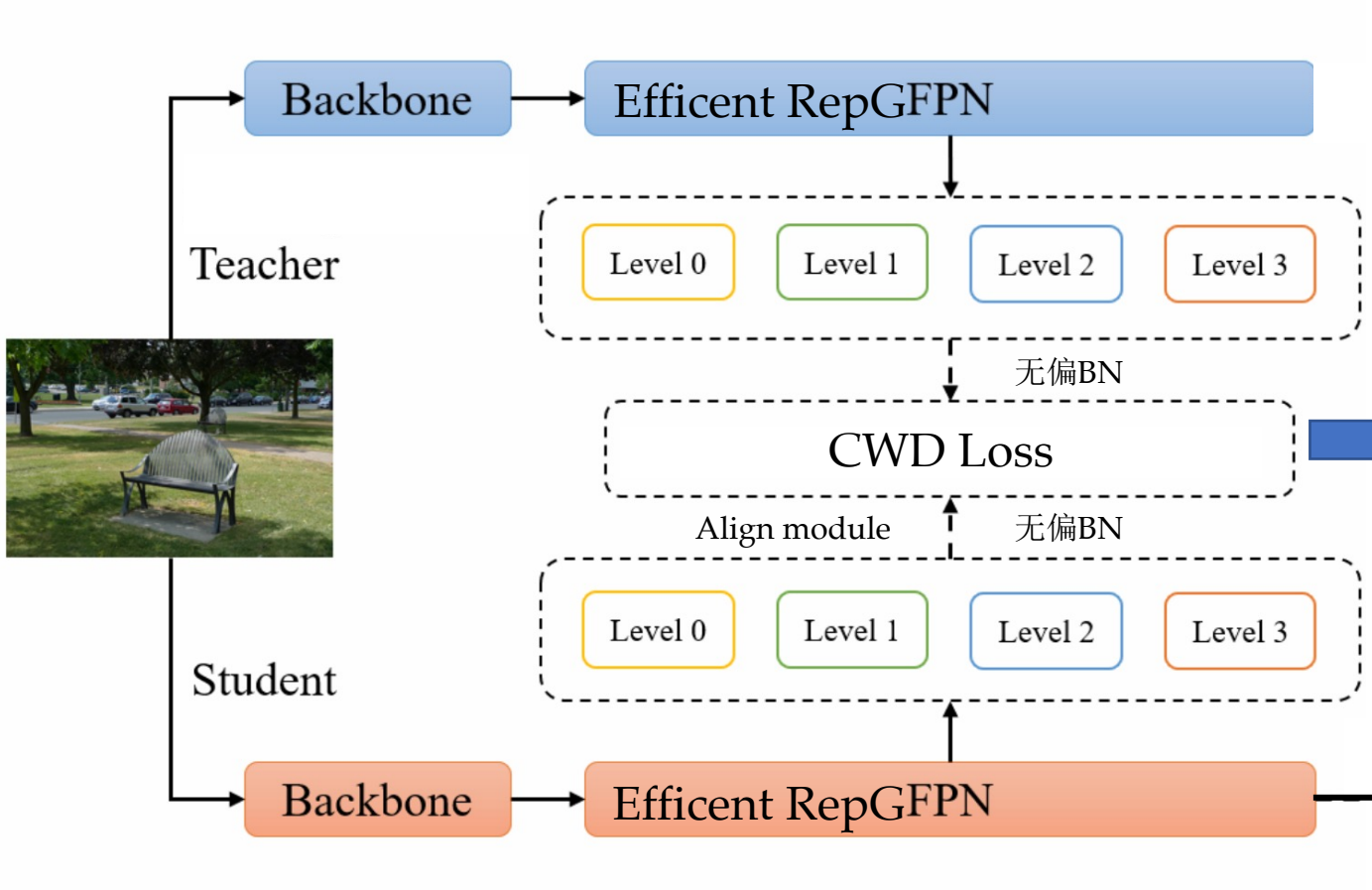
DAMO-YOLO基本原理

全尺度模型蒸馏

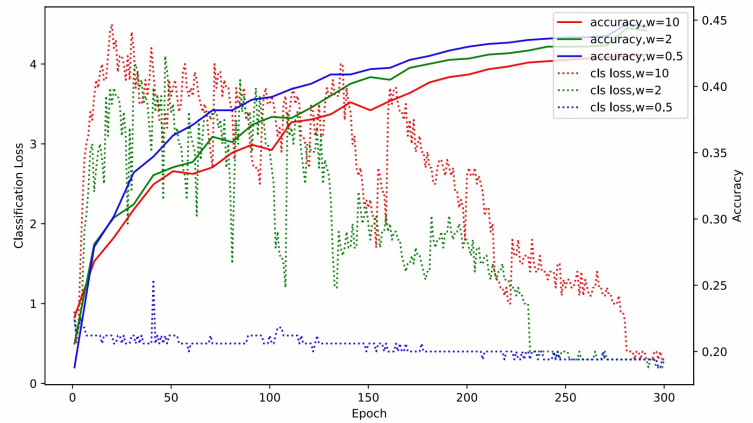
- 特征蒸馏+AlignModule+无偏BN = 全尺度模型蒸馏
- 调参free，支持全系列模型，支持异构蒸馏。
- 蒸馏链条L(CSP)->M(CSP)->S(Res)->T(Res)

全尺度模型蒸馏在T/S/M的效果

Method	Size	Latency(ms)	GFLOPs	Params(M)	AP
DAMO-YOLO-T	640	2.78	18.1	8.5	41.8
DAMO-YOLO-T*	640	2.78	18.1	8.5	43.0
DAMO-YOLO-S	640	3.83	37.8	16.3	45.6
DAMO-YOLO-S*	640	3.83	37.8	16.3	46.8
DAMO-YOLO-M	640	5.62	61.8	28.2	48.7
DAMO-YOLO-M*	640	5.62	61.8	28.2	50.0



DAMO-YOLO蒸馏框架



DAMO-YOLO-S蒸馏权重与分类损失的变化曲线

DAMO-YOLO基本原理

降本、
增效

DAMO-YOLO

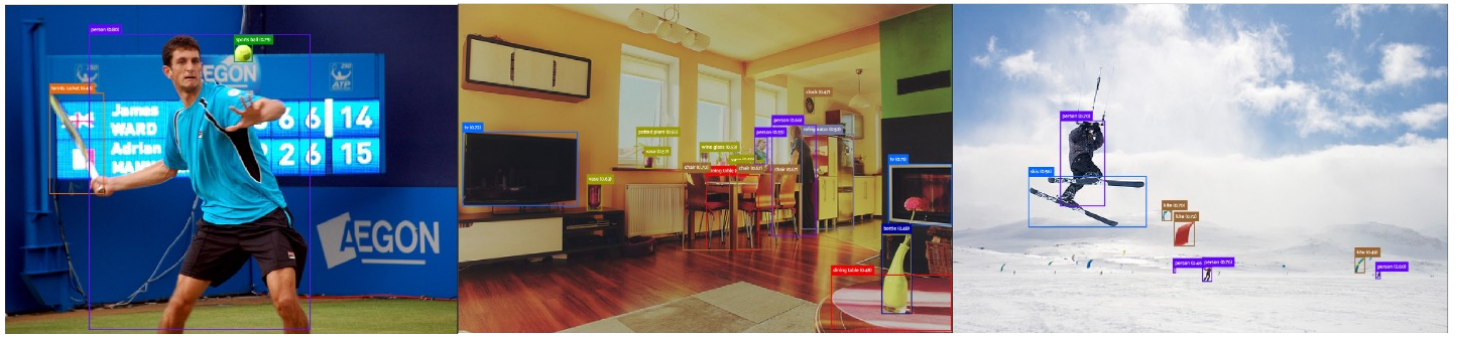
from Alibaba Group

MAE-NAS

低成本自定义模型

Efficient
RepGFPN

高效特征融合，提升多尺度检测能力



全尺度
蒸馏

无痛涨点，简便易用

HeavyNeck
范式

DAMO-YOLO模型已上线ModelScope，欢迎试用！

ModelScope：https://www.modelscope.cn/models/damo/cv_tinynas_object-detection_damoyolo/summary
Github：https://github.com/tinyvision/DAMO-YOLO
Arxiv：https://arxiv.org/abs/2211.15444

Alibaba Group 阿里巴巴集团

DAMO-YOLO

from Alibaba Group

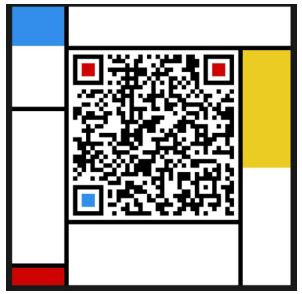
感谢聆听！

DAMO-YOLO已上线ModelScope，快去试试吧~

https://www.modelscope.cn/models/damo/cv_tinynas_object-detection_damoyolo/summary

实习生招聘中，欢迎投递到 xiuyu.sxy@alibaba-inc.com

钉钉



微信