标签分配与端到端目标检测

智源社区

MEGVII 旷视

主讲人：王剑锋

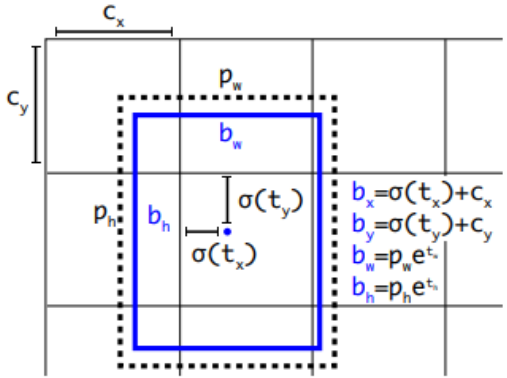
2021年2月4日

从anchor说起

MEGVII 旷视

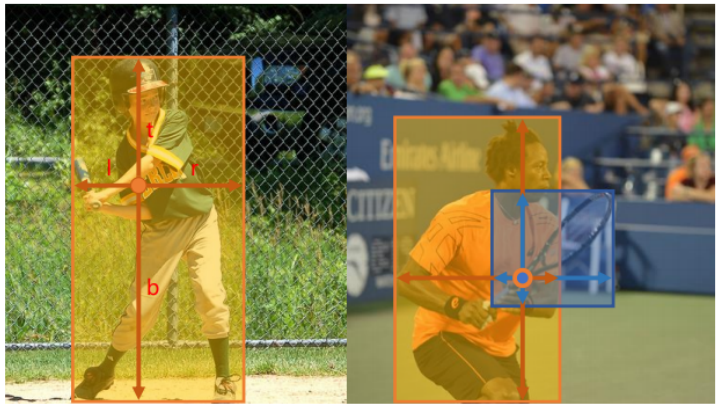
由于CNN的平移等变性，通常学习相对坐标

既然是相对坐标，就需要一个“零点”



当“零点”是**一个框**的时候，就是**anchor-based**方法，“零点”是**anchor box**

当“零点”是**一个点**的时候，就是**anchor-free**方法，“零点”实际上是**anchor point**



anchor-based / anchor-free某种意义上不是一个合适的划分方式

[1] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." CVPR. 2017.
[2] Tian, Zhi, et al. "Fcos: Fully convolutional one-stage object detection." ICCV. 2019.

Label Assignment

MEGVII 旷视

dense prediction的输出结果数远多于目标数，故而需要研究目标和输出结果的匹配问题，即“哪个（些）输出对应哪个目标”

举例

- RetinaNet根据**anchor和目标的IoU**确定正负样本
- FCOS根据**目标中心区域和目标的尺度**确定正负样本

两个要素： **尺度分配、空间分配**

Method	Prior	Instance		*AP*
		scale	spatial	
RetinaNet [10]	anchor	size & IoU	IoU	36.3
FreeAnchor [25]	anchor	size & IoU	top-k weighting, IoU	38.7
ATSS [24]	anchor	size & IoU	top-k, dynamic IoU	39.3
GuidedAnchoring [20]	dynamic anchor	size & IoU	IoU	37.1
FCOS* [19]	center	range	radius	38.7
FSAF [27]	anchor & center	loss	IoU & radius	37.2
AutoAssign (Ours)	**Center Weighting**	**Confidence Weighting**		**40.5**

[1] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." ICCV. 2017.
[2] Tian, Zhi, et al. "Fcos: Fully convolutional one-stage object detection." ICCV. 2019.

Label Assignment

虽然label assignment的表现形式非常多，但可以用一个比较统一的数学形式概括

N：ground-truth个数（目标个数），A：prediction个数（输出结果个数），通常 $N \ll A$

定义匹配矩阵：

$$M_{N\times A}, \quad s.t. \quad m_{ij} \in \{0,1\}, \sum_i \|m_{ij}\|_0 \leq 1$$

- 当 $m_{ij} \in \{0,1\}$ 时，$\sum_i \|m_{ij}\|_0 = \sum_i m_{ij} \leq 1$ 表示一个prediction至多对应一个ground-truth
- 反过来没有任何约束，通常是一对多的关系，即一个ground-truth对应多个prediction
- 非单射非满射：允许anchor不对应ground-truth，允许ground-truth没有anchor对应
- 在实践中，用1表示match，0表示mismatch，-1表示ignore

[1] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." ICCV. 2017.
[2] Tian, Zhi, et al. "Fcos: Fully convolutional one-stage object detection." ICCV. 2019.
[3] Zhu, Benjin, et al. "AutoAssign: Differentiable Label Assignment for Dense Object Detection." arXiv:2007.03496.

Label Assignment

RetinaNet:

$$\boldsymbol{Q}_{N\times A} = \mathrm{IoU}\left(\boldsymbol{gt}_i, \boldsymbol{anchor}_j\right)$$

$$\boldsymbol{M}_{N\times A} = \begin{cases} \mathbf{1}, & \boldsymbol{q}_{ij} > \mathbf{0.5} \\ \mathbf{0}, & \boldsymbol{q}_{ij} < \mathbf{0.4} \\ \mathbf{-1}, & \boldsymbol{others} \end{cases}$$

ATSS:

$$\boldsymbol{D}_{N\times A} = \mathrm{D}_{\mathrm{Euclidean}}\left(\mathrm{center}\left(\boldsymbol{gt}_i\right), \mathrm{center}\left(\boldsymbol{anchor}_j\right)\right)$$

$$\boldsymbol{Q}_{N\times A} = \mathrm{IoU}\left(\boldsymbol{gt}_i, \boldsymbol{anchor}_j\right)$$

$$\boldsymbol{M}_{N\times A} = \begin{cases} \mathbf{1}, & \boldsymbol{q}_{ij} > \mathrm{mean}\left(\boldsymbol{q}_{ij}\right) + \mathrm{std}\left(\boldsymbol{q}_{ij}\right) \quad \boldsymbol{s.t.} \quad \boldsymbol{j} \in \mathrm{topk}_{\boldsymbol{p}}\left(\boldsymbol{d}_{ij}\right) \\ \mathbf{0}, & \boldsymbol{others} \end{cases}$$

FCOS:

$$\boldsymbol{D}_{N\times A} = \mathrm{D}_{\mathrm{Chebyshev}}\left(\mathrm{center}\left(\boldsymbol{gt}_i\right), \mathrm{center}\left(\boldsymbol{anchor}_j\right)\right)$$

$$\boldsymbol{M}_{N\times A} = \begin{cases} \mathbf{1}, & \boldsymbol{d}_{ij} < \mathrm{radius} \wedge \max\left(\Delta\left(\boldsymbol{gt}_i, \mathrm{center}\left(\boldsymbol{anchor}_j\right)\right)\right) \in \left[\boldsymbol{s}_{\boldsymbol{p-1}}, \boldsymbol{s}_{\boldsymbol{p}}\right] \\ \mathbf{0}, & \boldsymbol{others} \end{cases}$$

[1] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." ICCV. 2017.
[2] Tian, Zhi, et al. "Fcos: Fully convolutional one-stage object detection." ICCV. 2019.
[3] Zhang, Shifeng, et al. "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection." CVPR. 2020.

Label Assignment

因而，常见的regression-based目标检测方法的loss function可以表示为：

$$\boldsymbol{L} = \frac{\mathbf{1}}{\sum \boldsymbol{m_{ij}}} \sum \left(\boldsymbol{m_{ij}} \times \mathrm{L}\left(\boldsymbol{gt_i}, \boldsymbol{pred_j} \right) \right), \qquad \boldsymbol{s.t.} \quad \boldsymbol{m_{ij}} \in \{\mathbf{0}, \mathbf{1}\}, \sum_{\boldsymbol{i}} \left\| \boldsymbol{m_{ij}} \right\|_{\mathbf{0}} \leq \mathbf{1}$$

部分工作将 $\boldsymbol{m_{ij}} \in \{\mathbf{0}, \mathbf{1}\}$ 松弛为 $\boldsymbol{m_{ij}} \in [\mathbf{0}, \mathbf{1}]$

例如：NoisyAnchor、IoU-Balanced Loss

Focal Loss的加权项 $\alpha(1-\boldsymbol{p})^{\gamma}$ 也可以视为一种 $\boldsymbol{m_{ij}}$

部分工作进一步松弛了 $\sum_i \left\| \boldsymbol{m_{ij}} \right\|_{\mathbf{0}} \leq \mathbf{1}$，允许一个prediction对应多个ground-truth，甚至正负样本loss拥有各自的 $\boldsymbol{M_{N \times A}}$

例如：FreeAnchor、AutoAssign

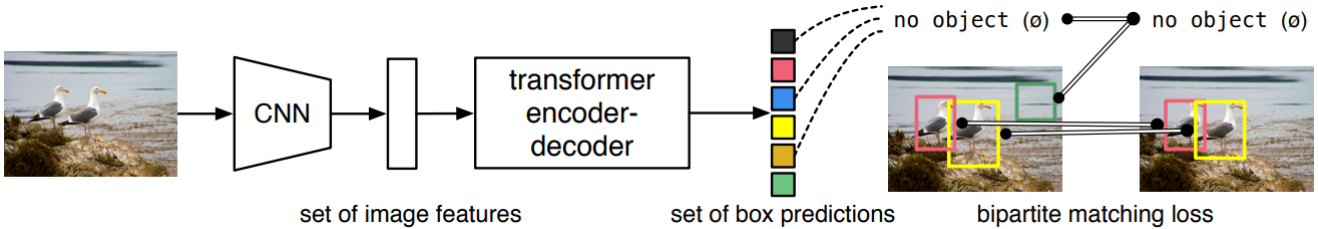
可以发现label assignment与loss function是有紧密联系的

[1] Li, Hengduo, et al. "Learning from noisy anchors for one-stage object detection." CVPR. 2020.
[2] Wu, Shengkai, and Xiaoping Li. "Iou-balanced loss functions for single-stage object detection." arXiv:1908.05641.
[3] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." ICCV. 2017.
[4] Zhang, Xiaosong, et al. "Freeanchor: Learning to match anchors for visual object detection." NeurIPS. 2019.
[5] Zhu, Benjin, et al. "AutoAssign: Differentiable Label Assignment for Dense Object Detection." arXiv:2007.03496.

NMS-free

DETR，基于Transformer结构，用bipartite matching完成targets与queries的匹配（可以视为广义上的label assignment）

端到端 (End-to-end)，**无NMS后处理**



去掉NMS的好处：

- 在遮挡、拥挤场景下会降低recall
- 难以并行，在某些场景下比较耗时（比如某些芯片上）
- 阻碍了某些下游任务的端到端化
- ……

有一个自然的问题：端到端检测必须依赖Transformer结构么？全卷积网络可以么？

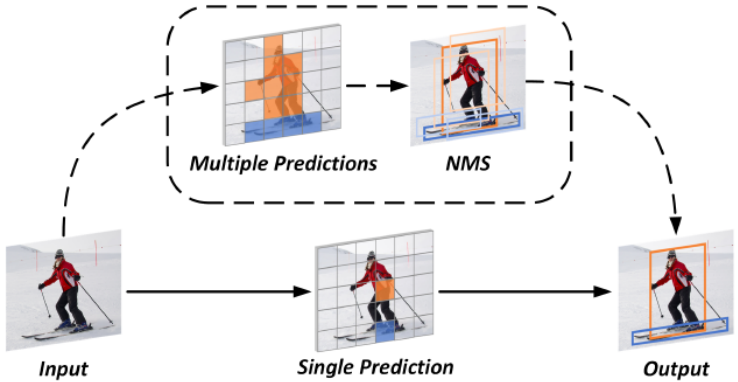
[1] Carion, Nicolas, et al. "End-to-end object detection with transformers." ECCV. 2020.

One-to-many v.s. One-to-one

常见的label assignment环节是one-to-many匹配，这导致网络必然需要一个NMS或NMS类似物做many-to-one去重。

MultiBox采用了one-to-one匹配

DETR ≈ 以Transformer为网络结构的MultiBox



我们用前面提到的数学形式表示：

$$\boldsymbol{L} = \frac{\mathbf{1}}{\sum \boldsymbol{m_{ij}}} \sum \left(\boldsymbol{m_{ij}} \times \mathrm{L}\left(\boldsymbol{gt_i}, \boldsymbol{pred_j} \right) \right), \qquad \boldsymbol{s.t.} \quad \boldsymbol{m_{ij}} \in \{\mathbf{0}, \mathbf{1}\}, \sum_{\boldsymbol{i}} \left\| \boldsymbol{m_{ij}} \right\|_{\mathbf{0}} \leq \mathbf{1}, \sum_{\boldsymbol{j}} \left\| \boldsymbol{m_{ij}} \right\|_{\mathbf{0}} = 1$$

倘若将其视为组合优化问题：

将 $\mathrm{L}\left(\boldsymbol{gt_i}, \boldsymbol{pred_j} \right)$ 视为定值，求解 $\boldsymbol{m_{ij}}$ 最小化 $\boldsymbol{L}$ ——等价于**以 L 为cost的bipartite matching**

[1] Erhan, Dumitru, et al. "Scalable object detection using deep neural networks." CVPR. 2014.
[2] Carion, Nicolas, et al. "End-to-end object detection with transformers." ECCV. 2020.
[3] Wang, Jianfeng, et al. "End-to-end object detection with fully convolutional network." arXiv:2012.03544.

One-to-many v.s. One-to-one

在dense prediction上我们能不能只依赖one-to-one label assignment，比较完美地去掉NMS？

Assignment	Rule	Method	mAP w/ NMS	mAP w/o NMS	mAP Δ	mAR w/ NMS	mAR w/o NMS	mAR Δ
One-to-many	Hand-designed	FCOS [42] baseline *	40.5	12.1	-28.4	58.3	52.8	-5.5
One-to-one	Hand-designed	Anchor	37.2	35.8	-1.4	57.0	59.2	+2.2
		Center	37.2	33.6	-3.6	57.8	59.7	+1.9
	Prediction-aware	Foreground loss	38.3	37.1	-1.2	58.6	**61.4**	**+2.8**
		POTO	38.6	38.0	-0.6	57.9	60.5	+2.6
		POTO+3DMF	40.0	39.8	-0.2	58.8	60.9	+2.1
Mixture **	Prediction-aware	POTO+3DMF+Aux	**41.2**	**41.1**	**-0.1**	**58.9**	61.2	+2.3

loss function和evaluation metrics往往并不一致，它常常要为优化问题做一些妥协（比如做一些加权等等）

也就是说，loss并不一定是bipartite matching的最佳cost

由此，我们定义了matching quality：

$$\boldsymbol{m_{ij}} = \mathbb{1}\left[\boldsymbol{j} \in \boldsymbol{\Omega_j}\right] \cdot \left[\boldsymbol{prob_j}(\boldsymbol{c_i})\right]^{1-\alpha} \cdot \left[\mathrm{IoU}\left(\boldsymbol{gt_i}, \boldsymbol{pred_j}\right)\right]^{\alpha}$$

[1] Tian, Zhi, et al. "Fcos: Fully convolutional one-stage object detection." ICCV. 2019.
[2] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." ICCV. 2017.
[3] Carion, Nicolas, et al. "End-to-end object detection with transformers." ECCV. 2020.
[4] Wang, Jianfeng, et al. "End-to-end object detection with fully convolutional network." arXiv:2012.03544.

One-to-many v.s. One-to-one

为什么是**加权几何平均数**：$[\boldsymbol{prob}]^{1-\alpha} \cdot [\boldsymbol{IoU}]^{\alpha}$ ，

而不是**加权算术平均数**：$(1-\alpha) \cdot [\boldsymbol{prob}] + \alpha \cdot [\boldsymbol{IoU}]$ ？（NoisyAnchor采用了类似公式）

假设检测loss由CE loss和IoU loss构成，

$$
\begin{aligned}
\boldsymbol{L_{det}} &= \boldsymbol{L_{cls}} + \boldsymbol{\lambda} \boldsymbol{L_{reg}} \\
&= -\log(\boldsymbol{prob}) - \boldsymbol{\lambda} \log(\boldsymbol{IoU}) \\
&= -\log\left(\boldsymbol{prob} \times \boldsymbol{IoU}^{\lambda}\right)
\end{aligned}
$$

从MLE形式可以看出 $\boldsymbol{prob}$ 和 $\boldsymbol{IoU}$ 是乘性关系

实验也证明了乘性关系效果更好

Table 4. The effect of various quality functions on COCO *val* set. '/' is used to distinguish between results without and with NMS. 'Add' and 'Mul' indicate two fusion functions.

Method	α	mAP	AP_{50}	AP_{75}
Add	0.2	36.0 / 36.2	**55.7** / 57.0	38.7 / 38.3
	0.5	37.3 / 37.8	54.9 / 57.4	40.5 / 40.4
	0.8	29.3 / 35.6	40.3 / 53.4	32.8 / 38.4
Mul	0.8	**38.0 / 38.6**	55.2 / **57.6**	**41.4 / 41.3**

[1] Li, Hengduo, et al. "Learning from noisy anchors for one-stage object detection." CVPR. 2020.
[2] Wang, Jianfeng, et al. "End-to-end object detection with fully convolutional network." arXiv:2012.03544.

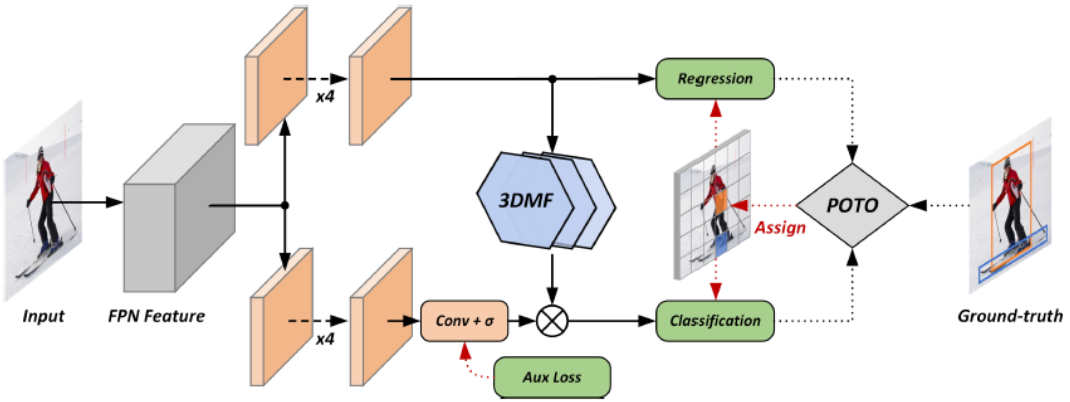
One-to-many v.s. One-to-one

one-to-one的劣势：

- 需要输出的feature更sharp，对网络表征能力要求较高
- 相较于one-to-many，监督更弱，收敛速度较慢

我们的改进尝试：

- 引入3D Max Filtering增强表征能力
- 引入one-to-many auxiliary loss增强监督



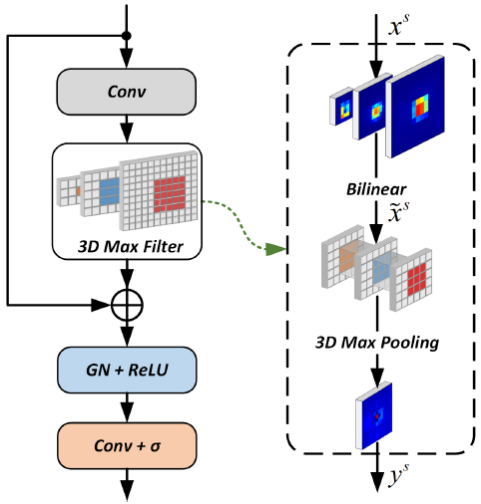
[1] Wang, Jianfeng, et al. "End-to-end object detection with fully convolutional network." arXiv:2012.03544.

One-to-many v.s. One-to-one

MEGVII 旷视

3D Max Filtering

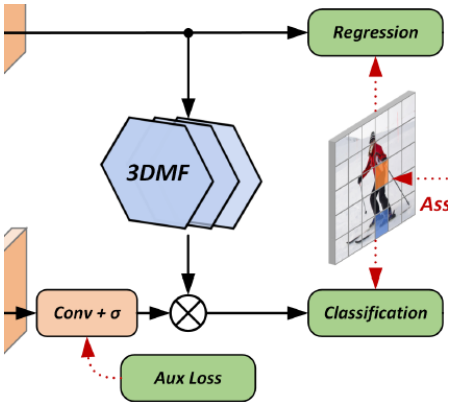
- 卷积是线性滤波器，学习max操作是比较困难的
- 希望引入非线性滤波器
- 希望引入跨层操作



one-to-many auxiliary loss

- auxiliary loss的具体assignment方法并不关键
- 乘法是auxiliary loss可以work的关键

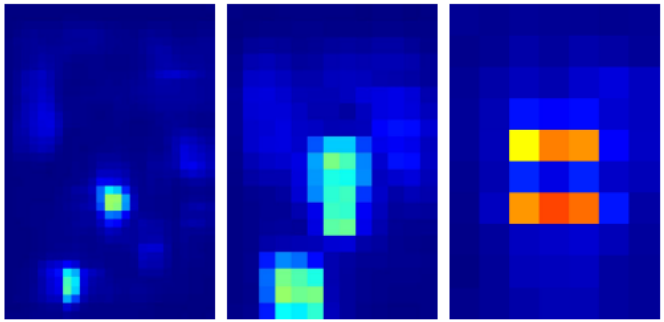
最近也有工作以one-to-many为主，one-to-one为辅



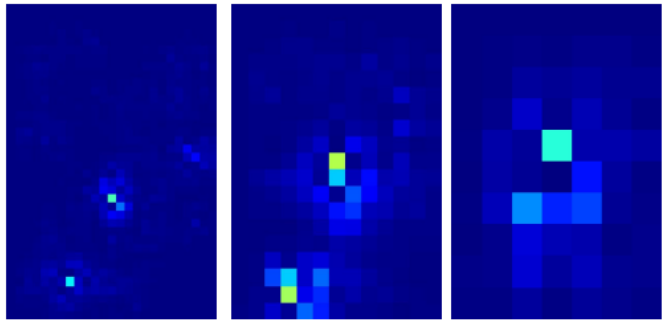
[1] Wang, Jianfeng, et al. "End-to-end object detection with fully convolutional network." arXiv:2012.03544.
[2] Zhou, Qiang, et al. "Object Detection Made Simpler by Eliminating Heuristic NMS." arXiv:2101.11782.

可视化

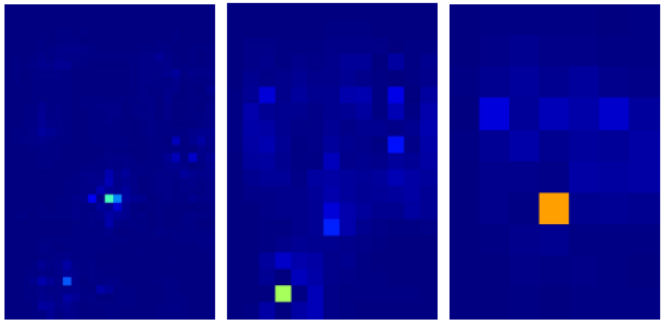
Image



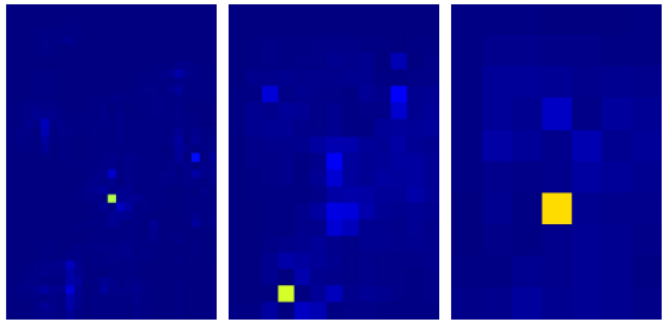
(a) FCOS baseline



(b) POTO



(c) POTO+3DMF



(d) POTO+3DMF+Aux

[1] Wang, Jianfeng, et al. "End-to-end object detection with fully convolutional network." arXiv:2012.03544.

实验结果

Table 3. Results of POTO with different configurations of α and spatial prior on COCO *val* set. $\alpha = 0$ is equivalent to considering classification alone, $\alpha = 1$ is equivalent to considering regression alone. 'center sampling' and 'inside box' both follow FCOS [42]. '/' is used to distinguish between results without and with NMS.

α	center sampling	inside box	global
0.0	33.5 / 33.6	24.1 / 24.2	1.9 / 2.1
0.2	33.7 / 33.9	28.8 / 28.8	19.4 / 19.5
0.4	35.0 / 35.2	32.7 / 32.8	28.3 / 28.4
0.6	36.6 / 36.9	35.3 / 35.5	34.7 / 34.9
0.8	**38.0 / 38.6**	37.4 / 37.9	37.3 / 37.9
1.0	11.8 / 29.7	4.5 / 13.0	non-convergence

- α 越低，分类权重越大，有无NMS的差距越小，但绝对性能也会降低；α 太高也不好，我们后续所有实验用 $\alpha = 0.8$
- 在 α 合理的情况下，空间先验不是必须的，但空间先验能够在匹配过程中帮助排除不好的区域，提升绝对性能

[1] Wang, Jianfeng, et al. "End-to-end object detection with fully convolutional network." arXiv:2012.03544.

实验结果

MEGVII 旷视

Table 7. The experiments of the proposed framework with larger backbone on COCO2017 *test-dev* set. The hyper-parameters of all models follow the official settings.

Backbone	Model	mAP	FPS
ResNet-101	RetinaNet [20]	40.7	13.6
	FCOS [42]	**43.2**	16.7
	Ours (w/o NMS)	42.9	15.9
ResNeXt-101+DCN	RetinaNet [20]	44.5	6.4
	FCOS [42]	46.5	6.8
	Ours (w/o NMS)	**47.6**	6.6

Table 8. The comparison of fully convolutional detectors on CrowdHuman *val* set. All models are based on the ResNet-50 backbone. 'Aux' indicates the auxiliary loss.

Method	AP_{50}	mMR	Recall	FPS
RetinaNet [20]	81.7	57.6	88.6	16.5
FCOS [42]	86.1	55.2	94.3	22.8
ATSS [46]	87.1	50.1	94.0	22.4
Ground-truth (w/ NMS)	-	-	95.1	-
POTO	88.7	52.0	96.6	23.1
POTO+3DMF	89.0	50.3	96.4	21.3
POTO+3DMF+Aux	**89.2**	**49.6**	**96.6**	21.3

CrowdHuman的ground-truth做NMS threshold=0.6，只有95.1%的Recall，这也是NMS方法的理论上限

更多结果更新在https://github.com/Megvii-BaseDetection/DeFCN

[1] Wang, Jianfeng, et al. "End-to-end object detection with fully convolutional network." arXiv:2012.03544.

用人工智能造福大众

MEGVII 旷视