



## 一种用于目标跟踪边界框回归的光滑IoU损失

李功 赵巍 刘鹏 唐降龙

### Smooth-IoU Loss for Bounding Box Regression in Visual Tracking

LI Gong, ZHAO Wei, LIU Peng, TANG Xiang-Long

在线阅读 View online: <https://doi.org/10.16383/j.aas.c210525>

---

## 您可能感兴趣的其他文章

### 传感器饱和的非线性网络化系统模糊滤波

Fuzzy Filtering for Nonlinear Networked Systems Subject to Sensor Saturations

自动化学报. 2021, 47(5): 1149–1158 <https://doi.org/10.16383/j.aas.c180778>

### L2损失大规模线性非平行支持向量顺序回归模型

L2-loss Large-scale Linear Nonparallel Support Vector Ordinal Regression

自动化学报. 2019, 45(3): 505–517 <https://doi.org/10.16383/j.aas.2018.c170438>

### 基于稀疏表示的视频目标跟踪研究综述

Research of Sparse Representation-based Visual Object Tracking: A Survey

自动化学报. 2018, 44(10): 1747–1763 <https://doi.org/10.16383/j.aas.2018.c170209>

### 一种高斯渐进滤波框架下的目标跟踪方法

A Target Tracking Method in Gaussian Progressive Filtering Framework

自动化学报. 2018, 44(12): 2250–2258 <https://doi.org/10.16383/j.aas.2018.c170421>

### 基于相关滤波器的视频跟踪方法研究进展

Research Progress of Visual Tracking Methods Based on Correlation Filter

自动化学报. 2019, 45(2): 265–275 <https://doi.org/10.16383/j.aas.2018.c170394>

### 融合显著性与运动信息的相关滤波跟踪算法

Correlation Filter Based Visual Tracking Integrating Saliency and Motion Cues

自动化学报. 2021, 47(7): 1572–1588 <https://doi.org/10.16383/j.aas.c190122>

# 一种用于目标跟踪边界框回归的光滑 IoU 损失

李功<sup>1</sup> 赵巍<sup>1</sup> 刘鹏<sup>1</sup> 唐降龙<sup>1</sup>

**摘要** 边界框回归分支是深度目标跟踪器的关键模块, 其性能直接影响跟踪器的精度. 评价精度的指标之一是交并比 (Intersection over union, IoU). 基于 IoU 的损失函数取代了  $\ell_n$ -norm 损失成为目前主流的边界框回归损失函数, 然而 IoU 损失函数存在 2 个固有缺陷: 1) 当预测框与真值框不相交时 IoU 为常量 0, 无法梯度下降更新边界框的参数; 2) 在 IoU 取得最优值时其梯度不存在, 边界框很难收敛到 IoU 最优处. 揭示了在回归过程中 IoU 最优的边界框各参数之间蕴含的定量关系, 指出在边界框中心处于特定位置时存在多种尺寸不同的边界框使 IoU 损失最优的情况, 这增加了边界框尺寸回归的不确定性. 从优化两个统计分布之间散度的视角看待边界框回归问题, 提出了光滑 IoU (Smooth-IoU, SIoU) 损失, 即构造了在全局上光滑 (即连续可微) 且极值唯一的损失函数, 该损失函数自然蕴含边界框各参数之间特定的最优关系, 其唯一取极值的边界框可使 IoU 达到最优. 光滑性确保了在全局上梯度存在使得边界框更容易回归到极值处, 而极值唯一确保了在全局上可梯度下降更新参数, 从而避开了 IoU 损失的固有缺陷. 提出的光滑损失可以很容易取代 IoU 损失集成到现有的深度目标跟踪器上训练边界框回归, 在 LaSOT、GOT-10k、TrackingNet、OTB2015 和 VOT2018 测试基准上所取得的结果, 验证了光滑 IoU 损失的易用性和有效性.

**关键词** 光滑 IoU 损失,  $\ell_n$ -norm 损失, 边界框回归, 目标跟踪

**引用格式** 李功, 赵巍, 刘鹏, 唐降龙. 一种用于目标跟踪边界框回归的光滑 IoU 损失. 自动化学报, 2023, 49(2): 288–306

**DOI** 10.16383/j.aas.c210525

## Smooth-IoU Loss for Bounding Box Regression in Visual Tracking

LI Gong<sup>1</sup> ZHAO Wei<sup>1</sup> LIU Peng<sup>1</sup> TANG Xiang-Long<sup>1</sup>

**Abstract** The branch of bounding box regression is a critical module in visual object trackers, and its performance directly affects accuracy of a tracker. One of evaluation metrics used to measure accuracy is intersection over union (IoU). The IoU loss which was proposed to replace  $\ell_n$ -norm loss for bounding box regression is increasingly popular. However, there are two inherent issues in IoU loss: One is that the parameters of bounding box can not be updated via gradient descent if the predicted box does not intersect with ground-truth box; the other is the gradient of the optimal IoU does not exist, so it is difficult to make the predicted box regressed to the IoU optimum. We reveal the explicit relationship among the parameters of IoU optimal bounding box in regression process, and point out that the size of a predicted box which makes IoU loss optimal is not unique when its center is in specific areas, increasing the uncertainty of bounding box regression. From the perspective of optimizing divergence between two distributions, we propose a smooth-IoU (SIoU) loss, which is a globally smooth (continuously differentiable) loss function with unique extremum. The smooth-IoU loss naturally implicates a specific optimal relationship among the parameters of bounding box, and its gradient over the global domain exists, making it easier to regress the predicted box to the extremal bounding box, and the unique extremum ensures that the parameters can be updated via gradient descent. In addition, the proposed smooth-IoU loss can be easily incorporated into existing trackers by replacing the IoU-based loss to train bounding box regression. Extensive experiments on visual tracking benchmarks including LaSOT, GOT-10k, TrackingNet, OTB2015, and VOT2018 demonstrate that smooth-IoU loss achieves state-of-the-art performance, confirming its effectiveness and efficiency.

**Key words** Smooth-IoU loss,  $\ell_n$ -norm loss, bounding box regression, visual tracking

**Citation** Li Gong, Zhao Wei, Liu Peng, Tang Xiang-Long. Smooth-IoU loss for bounding box regression in visual tracking. *Acta Automatica Sinica*, 2023, 49(2): 288–306

收稿日期 2021-06-11 录用日期 2021-09-17

Manuscript received June 11, 2021; accepted September 17, 2021

国家自然科学基金 (51935005), 基础科研项目 (JCKY20200603 C010), 空间智能控制技术重点实验室基金 (ZDSYS-2018-02) 资助

Supported by National Natural Science Foundation of China (51935005), Basic Scientific Research Projects (JCKY20200603C010), and Mutual Fund of Space Intelligent Control Technology Key Laboratory (ZDSYS-2018-02)

目标跟踪是计算机视觉领域里的基础任务之一. 随着深度学习在各个领域里日益成熟的广泛应用, 基于深度网络的目标跟踪方法<sup>[1]</sup>取得了显著的

本文责任编辑 桑农

Recommended by Associate Editor SANG Nong

1. 哈尔滨工业大学模式识别与智能系统研究中心 哈尔滨 150001  
1. Pattern Recognition and Intelligence System Research Center, Harbin Institute of Technology, Harbin 150001

提升和进步. 与目标检测方法<sup>[2]</sup>类似, 边界框预测模块在目标跟踪方法里也是至关重要的一环, 它的性能直接影响目标跟踪模型准确性. 交并比 (Intersection over union, IoU) 是衡量跟踪准确性的重要评估指标, 定义为  $\text{IoU}(\tilde{B}; B) = |\tilde{B} \cap B| / |\tilde{B} \cup B|$  用来衡量预测的边界框  $\tilde{B}$  与真值框  $B$  的相近程度. 对于两个不同的跟踪器, 即便跟踪器的分类模块都能够定位到目标所在位置, 但边界框预测模块的性能不同仍可能导致预测结果的 IoU 相差甚远, 所以训练边界框回归准确甚为重要. 从时间发展的顺序上看, 边界框回归方法可分为 2 类: 第 1 类是基于  $\ell_n$ -norm 损失的回归, 可表示为  $\ell_n(\tilde{B}; B) = \|\tilde{B} - B\|_n$ . 其中常用的两种损失  $\ell_1$ -norm 和  $\ell_2$ -norm 都有缺陷, 前者难以收敛到更高的精度, 而后者在训练初始时不稳定. 虽然 Girshick<sup>[3]</sup> 提出的  $\ell_1$ -smooth 损失, 可以较好地解决上述两个问题, 但是基于  $\ell_n$ -norm 的回归损失更备受诟病的是边界框各个参数在优化过程中相互独立, 缺乏对 IoU 的考虑. 第 2 类损失函数是基于 IoU 损失的回归. IoU 损失<sup>[4]</sup>  $\mathcal{L}_{\text{IoU}}(\tilde{B}; B) = 1 - \text{IoU}(\tilde{B}; B)$  衍生自 IoU 指标, 避免预测框  $\tilde{B}$  的参数在回归过程中互不关联. 然而 IoU 损失  $\mathcal{L}_{\text{IoU}}$  有两个固有缺陷: 一个是当预测框与真值框不相交时  $\mathcal{L}_{\text{IoU}}$  为常量 1, 其梯度无法下降, 从而边界框回归分支的参数得不到更新; 另一个是在 IoU 取得最优值时  $\mathcal{L}_{\text{IoU}}$  的梯度不存在, 边界框难以收敛到 IoU 最优处. 其实  $\mathcal{L}_{\text{IoU}}$  的固有缺陷继承自 IoU 指标. 虽然随后的 GIoU (Generalized IoU) 损失<sup>[5]</sup>、DIoU (Distance-IoU) 损失<sup>[6]</sup>、CIoU (Complete-IoU) 损失<sup>[6]</sup> 和 EIoU (Efficient-IoU) 损失<sup>[7]</sup> 等对预测框  $\tilde{B}$  的中心或尺寸提出了不同的惩罚项来增加  $\mathcal{L}_{\text{IoU}}$  在边界框不交叠时的梯度, 但是附加的惩罚项并不能改善  $\mathcal{L}_{\text{IoU}}$  在最优值处的梯度不存在的问题.

上述基于 IoU 的损失方法<sup>[4-7]</sup> 已经暗示在回归过程中不应该忽视边界框参数之间的关系. 但是都没有明确究竟是何种关系. 边界框通常由 4 个参数确定, 但在不同的研究中参数的含义有所不同, 可以由边界框的中心和尺寸表示为  $B(x, y, w, h)$ <sup>[8-10]</sup>, 或者是由左上角点和右下角点表示为  $B(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ <sup>[11-12]</sup>, 又或是由给定的一点到四边的距离表示为  $B(x_t, x_b, x_l, x_r)$ <sup>[4, 13-14]</sup>. 其实上述表示都是等价的, 可以相互转化. 为方便下文描述, 本文统一以  $B(x, y; w, h)$  形式表示边界框. 不妨将预测框  $\tilde{B}(x, y; w, h)$  的 4 个参数划分为 2 组, 一组是中心位置  $(x, y)$ , 另一组是尺寸  $(w, h)$ . 显然, 对中心位置来说, 追求预测框中心与真值框中心重合总是最优

的, 即便有时预测框中心在某个邻域内波动不会使 IoU 下降. 一个显而易见的事实是, 不存在中心偏离可以使 IoU 上升的情况. 但对尺寸来说并非如此, 当预测框中心发生偏离时, 追求预测框的尺寸与真值框的尺寸相同却不是最优的.

本文明确给出在回归过程中边界框若取得 IoU 最优其参数之间应服从的定量关系. 概括地说, 中心  $(x, y)$  在回归过程不需要顾及此时尺寸  $(w, h)$  处于何种情况, 而尺寸  $(w, h)$  在回归过程中需要考虑到中心  $(x, y)$  所在何处, 最优尺寸  $(w^*, h^*) = \arg \min_{w, h} \text{IoU}(\tilde{B}; B) = f(x, y)$  与中心  $(x, y)$  存在明确的函数关系. 本文从一个新的角度看待边界框回归问题, 将边界框与二元统计分布作一一映射, 从优化两个统计分布之间散度的角度研究边界框回归. 散度量化了两个不同分布之间的差异, 这种散度自然蕴含预测框各参数之间的关系, 可以避免人为设计额外的惩罚项对预测框尺寸或形状做出限制. 本文从优化两个分布之间散度的角度提出了一种光滑 IoU (Smooth-IoU, SIoU) 损失, 该损失函数在全局上光滑, 对于不同的研究对象, 光滑的含义也有所区别. 在本文中称在开集  $X \in \mathbf{R}^n$  上的函数  $f: X \rightarrow \mathbf{R}$  是光滑的, 如果  $f$  是  $C^1$  类的,  $C^1$  类的函数必然是可微的. 在本文的定义下, 光滑性也可以称作连续可微性, 且极值唯一. 光滑性确保了在全局上梯度存在使得边界框更容易回归到极值处, 而极值唯一确保了在全局上可梯度下降更新参数, 从而克服了 IoU 损失的固有缺陷. 提出的光滑 IoU 损失自然蕴含边界框各参数之间特定的最优关系, 其唯一取极值的边界框可使 IoU 达到最优. 而且提出的光滑 IoU 损失具有比 IoU 损失更快的回归性能. 另外, 提出的光滑 IoU 损失可以很容易集成到具有边界框回归分支的视觉跟踪方法中. 为了评估提出的光滑 IoU 损失, 本文将其集成到跟踪深度网络模型 SiamFC++ 等中, 并在主要的基准 LaSOT、GOT10-k、TrackingNet、OTB2015 和 VOT2018 中进行了评估. 本文主要贡献为:

1) 明确给出在回归过程中最优边界框各参数之间满足的定量关系.

2) 从优化散度的角度提出光滑 IoU 损失, 该损失函数自然蕴含边界框各参数之间特定的最优关系, 在全局上连续可微, 且唯一极值可使 IoU 最优. 提出的损失函数避免了 IoU 损失的固有缺陷.

3) 提出的光滑 IoU 损失可以容易集成到先进跟踪网络方法, 在主流的测试基准 LaSOT、GOT-10k 和 TrackingNet 等上取得显著的提升.

## 1 相关工作

### 1.1 边界框回归损失

自 Fast-RCNN<sup>[3]</sup> 提出以来,  $l_1$ -smooth 损失<sup>[3]</sup> 就被广泛地应用在目标检测或跟踪任务中训练目标边界框的回归<sup>[8-10, 12]</sup>.  $l_1$ -smooth 损失结合了  $l_1$ -norm 和  $l_2$ -norm 中互补的良好性质. 然而, 对于相同的  $l_1$  或  $l_2$  误差 (只要不为 0), 可以回归出多种大小及形状不同的边界框, 而这些预测的边界框所对应的与真值框的 IoU 却不尽相同, 方差较大, 有较强的随机性, 不能准确地反映 IoU 指标. Rezaatofghi 等<sup>[5]</sup> 展示了一些  $l_1$ -norm 和  $l_2$ -norm 相同但 IoU 指标不同的示例. 为此, Yu 等<sup>[4]</sup> 将 IoU 指标演化为 IoU 损失  $\mathcal{L}_{IoU}$ , 通过直接优化 IoU 的方式边界框各参数可以作为一个整体进行回归. Rezaatofghi 等<sup>[5]</sup> 提出了一种广义的 IoU 指标 GIoU 及其演化的损失函数  $\mathcal{L}_{GIoU}$ , 以代替 IoU 用于评估和训练边界框回归, GIoU 损失纠正了 IoU 损失在预测框与真值框不相交时梯度无法下降的弊端. Zheng 等<sup>[6]</sup> 提出了 DIoU 损失函数  $\mathcal{L}_{DIoU}$ , 在 IoU 损失的基础上附加了一种关于预测框中心与真值框中心的归一化距离的惩罚项  $\mathcal{R}_D$ , 相较于 IoU 损失和 GIoU 损失加快了收敛速度. 同时, Zheng 等<sup>[6]</sup> 在 DIoU 的基础上发展出了 CIoU 损失函数  $\mathcal{L}_{CIoU}$ , 该损失函数综合考虑了 3 种几何属性, 分别是 IoU、中心点距离和宽高比率, 使得收敛速度进一步加快. 与 CIoU 类似, Zhang 等<sup>[7]</sup> 提出了另一种高效的 IoU 损失函数 EIoU 损失  $\mathcal{L}_{EIoU}$ , 该损失函数同样有 3 种几何因素的度量, 分别是 IoU、中心点的距离以及边长差异. 基于 IoU 损失可被统一地表示为  $\mathcal{L}_{*IoU}(\tilde{B}; B) = \mathcal{L}_{IoU}(\tilde{B}; B) + \mathcal{R}_*(\tilde{B}; B)$ , 其中  $\mathcal{R}_*(\tilde{B}; B)$  为各自不同的惩罚项. 本文提出的光滑 IoU 损失直接从散度方面构造全局光滑且极值唯一的损失函数, 没有以  $\mathcal{L}_{IoU}$  作为基本损失, 从而避免了  $\mathcal{L}_{IoU}$  带来的缺陷.

上述相关的边界框回归损失方法均假定边界框参数为确定变量而直接回归边界框, 除了这种处理方式外, 另一种处理方式是将描述边界框的关键点视为随机变量. 关键点可以是角点或中心点等, 通过预测关键点的热力图推断关键点最可能的位置. 热力图可视为关键点服从某种统计分布的假设, 对热力图的分布则通常采用 Focal loss<sup>[15]</sup> 训练. Gidaris 等<sup>[13]</sup> 在提出的 LocNet 中, 预测边界框的 4 个边框所在位置的置信度, 从而置信度最高的位置被推定为存在边框. Law 等<sup>[12]</sup> 在提出的 CornerNet 中设计了一种角点池化操作生成边界框的左上

角和右下角点的热力图, 并提出改进的 Focal loss 用于训练, 而对于角点精度的偏置则采用  $l_1$ -smooth 损失进行回归. 然而, 如果仅仅依靠预测左上角和右下角来确定边界框, 则容易导致错误匹配而误检. Zhou 等<sup>[14]</sup> 在提出的 CenterNet 中定义了一种适配的 Focal loss 用于训练边界框中心的热力图以减少错误匹配, 而对于中心点精度的偏置以及边界框尺寸则采用  $l_1$ -norm 损失进行回归. 另外, 文献 [11] 提出的方法则假设预测框的角点位置服从参数待学习的正态分布, 并假设真值框的角点位置服从狄拉克  $\delta$  分布. 通过以狄拉克  $\delta$  分布为目标优化正态分布实现边界框回归. 然而, 在上述文献 [11-14] 里用来描述边界框的关键点是独立优化的, 没有考虑关键点与 IoU 的关系, 其缺点与  $l_n$ -norm 在某种程度上类似, 都依赖于各关键点是否被预测得非常准确; 而且对热力图的训练增加了网络参数的数量和网络结构的复杂性.

本文提出的光滑 IoU 损失将表示边界框的 4 个参数视为一个整体进行回归, 在回归过程中能够照顾到 IoU 信息而产生 IoU 友好的结果, 而且本文提出的光滑 IoU 损失本质上是在最小化两个统计分布之间的散度, 在不增加网络复杂度的同时隐含地表达了将边界框关键点视为服从某种分布的随机变量这一处理方式.

### 1.2 深度跟踪器的边界框预测

GOTURN<sup>[16]</sup> 是第一个基于边界框回归的深度网络跟踪方法, 直接回归当前帧的目标框相对前一帧目标框的偏移. 随后的 SiamRPN<sup>[8]</sup> 和增强版的 DaSiamRPN<sup>[10]</sup> 结合了 SiamFC<sup>[17]</sup> 的孪生网络和 Fast R-CNN<sup>[3]</sup> 的区域候选网络 (Region proposal network, RPN), 估计边界框相对各个阳性锚框的偏移量, 并从中选出分类置信度最高的作为预测框. 然而 Jiang 等<sup>[18]</sup> 论证了分类置信度最高的边界框并不一定是与真值框吻合最优的. 因此, SPM-Tracker<sup>[19]</sup> 扩展了 SiamRPN 方法, 提出了精细匹配阶段, 旨在从粗略匹配阶段选出分数最高的  $k$  个候选框提炼最终预测框. 而 SiamRPN++<sup>[9]</sup> 则在分类置信度分支和边界框回归分支里提出了逐通道的互相关层, 并通过多层级联的方式提高了分类置信度和回归精度的正相关性. 上述基于锚框的深度网络方法通常采用  $l_1$ -smooth 损失训练边界框回归分支.

尽管基于锚框的跟踪方法仍有进一步优化网络和提升性能的空间, 但基于无锚框的跟踪方法则受到越来越多的青睐. 现有的研究工作和实验已经表明一些基于无锚框的深度网络方法比基于锚框的网络方法更准确, 同时网络参数的精简使得跟踪器在

训练和跟踪时更高效. SiamFC++<sup>[20]</sup> 建议目标跟踪模型的训练不应该介入尺度或长宽比率等先验分布的信息例如锚框, 其原因是定位和尺度等粗糙的锚框带来的误差可能拖累跟踪器的性能. SiamFC++ 摒弃了预设的锚框, 并将预测的目标从阳性锚框的偏移量转化为更精细的每个阳性位置到 4 条边线的距离. 随后的基于无锚框的方法如 SiamBAN<sup>[21]</sup>、SiamCAR<sup>[22]</sup>、Ocean<sup>[23]</sup> 和无锚框全卷积孪生跟踪器 (Anchor-free fully convolutional siamese tracker, AFST)<sup>[24]</sup> 等也采用类似的预测每个正样本位置到四边距离作为网络输出的方法. 值得一提的是, SiamBAN<sup>[21]</sup>、SiamCAR<sup>[22]</sup>、Ocean<sup>[23]</sup> 和 AFST<sup>[24]</sup> 均采用 IoU 损失训练边界框的回归.

本文将提出的光滑 IoU 损失, 应用到具有代表性的无锚框深度跟踪器 SiamFC++<sup>[20]</sup>、SiamBAN<sup>[21]</sup> 和 SiamCAR<sup>[22]</sup>, 通过替换其原有的 IoU 损失, 作为对比以评估光滑 IoU 损失的性能.

## 2 光滑 IoU 损失

### 2.1 从散度视角优化边界框 IoU 的动机

为方便描述, 先定义一些必要的表示记号.  $B_g(x_g, y_g; w_g, h_g)$  代表真值框,  $B_p(x_p, y_p; w_p, h_p)$  代表预测框,  $(x_\Delta, y_\Delta) := (x_p - x_g, y_p - y_g)$  代表预测框的中心位置相对于真值框中心的偏差.

图 1 给出了深度目标跟踪模型的基本框架, 本文不妨忽略与研究内容无关的分类或中心度分支  $\psi_{cls}$  的网络结果, 仅关注由孪生骨干网络  $\phi$  提取的

特征图经过边界框回归分支  $\psi_{reg}$  输出每一帧预测的目标边界框. 在训练深度目标跟踪模型的边界框回归分支时, 如果中心偏差  $(x_\Delta, y_\Delta)$  难以消除, 预测框的尺寸  $(w_p, h_p)$  若以  $\ell_2$ -norm 损失仍然向着真值框的尺寸  $(w_g, h_g)$  回归则不是 IoU 最优的, 而以 IoU 损失回归则在其最优的预测框尺寸  $(w_p^*, h_p^*)$  上又是不可微的. 所以一个自然的问题是如何即时调整预测框尺寸  $(w_p, h_p)$  的回归目标, 使损失函数面向更高的 IoU 指标光滑地回归.

为了解决上述问题, 本文从最小化统计分布之间散度的角度看待边界框回归问题. 首先本文将边界框与二元正态分布建立一一对应关系, 如图 2 所示. 具体地, 将预测框  $B_p(x_p, y_p; w_p, h_p)$  的中心位置  $(x_p, y_p)$  和尺寸  $(w_p, h_p)$  分别视为二元正态分布  $N(\mu_p, \Sigma_p)$  的均值  $\mu_p = (x_p, y_p)^T$  和边缘分布的标准差, 即  $\Sigma_p = \begin{bmatrix} w_p^2 & 0 \\ 0 & h_p^2 \end{bmatrix}$ , 这样预测框  $B_p(x_p, y_p; w_p, h_p)$  与二元正态分布  $N(\mu_p, \Sigma_p)$  建立了一一映射. 类似地, 真值框  $B_g(x_g, y_g; w_g, h_g)$  映射为均值为  $\mu_g = (x_g, y_g)^T$ , 协方差矩阵为  $\Sigma_g = \begin{bmatrix} w_g^2 & 0 \\ 0 & h_g^2 \end{bmatrix}$  的二元正态分布  $N(\mu_g, \Sigma_g)$ .

需要阐明的是, 与现有的相关工作<sup>[11-14]</sup> 的区别在于, 本文并不是假定边界框的 4 个参数本身为服从二元正态分布的随机变量, 而是将其一一映射为确定二元正态分布具体形式的参量, 可以理解为边界框蕴含了一种图像区域每个像素属于目标物体的置信分布, 该置信分布应该反映出越靠近边界框中

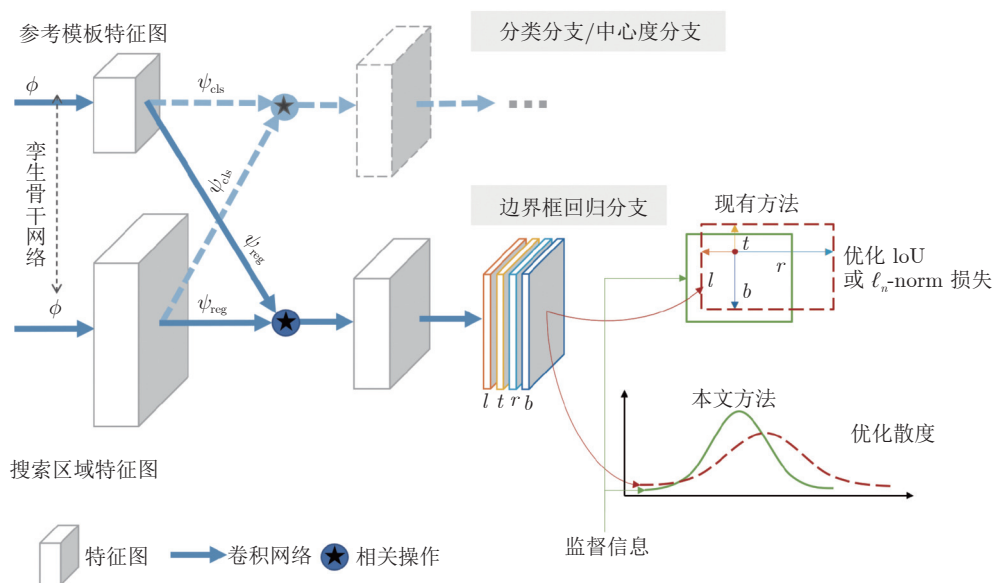


图 1 深度目标跟踪模型的边界框回归示意图

Fig.1 The schematic of bounding box regression in deep tracking model

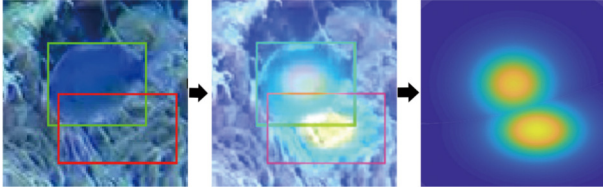


图 2 边界框类比为正态分布的示意图

Fig.2 The schematic of bounding box analogized as Gaussian distribution

心位置的像素属于目标物体的置信度越高的特点,从而隐含地表达了一种以边界框中心位置为关键点的热力图。

由此,将边界框的回归问题转化为最小化二元正态分布之间差异问题.以常见的 KL (Kullback-Leibler) 散度  $D_{KL}(N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)||N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p))$  量化二元正态分布  $N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$  和  $N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$  之间差异为例进行分析,基于 KL 散度的边界框回归损失函数为:

$$\mathcal{L}_{\text{reg}}^{KL}(B_p; B_g) \propto D_{KL}(N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)||N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)) = \frac{x_{\Delta}^2 + w_p^2}{2w_p^2} + \frac{y_{\Delta}^2 + h_g^2}{2h_p^2} - \ln \frac{w_g h_g}{w_p h_p} - 1 \quad (1)$$

显然,不同于 IoU 损失,式 (1) 在边界框全局上是可微的,而且式 (1) 表达了 KL 散度与 IoU 指标呈某种非线性负相关关系.当 KL 散度越小,说明 2 个二元正态分布  $N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$  和  $N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$  越接近,则与之对应的预测框  $B_p$  与真值框  $B_g$  就越接近,  $\text{IoU}(B_p; B_g)$  总体表现为上升趋势.当且仅当预测的边界框与真值框完全重合时, KL 散度减小到最小值 0,此时 IoU 提高到最大值 1.此外,式 (1) 里预测框各参数不是独立的回归.预测框的最优尺寸  $(w_p^*, h_p^*)$  与其中心偏差  $(x_{\Delta}, y_{\Delta})$  有关.当且仅当中心偏差  $(x_{\Delta}, y_{\Delta}) = (0, 0)$  时,其最优预测框尺寸  $(w_p^*, h_p^*)$  为真值框尺寸  $(w_g, h_g)$ ; 否则,相较于  $\ell_n$ -norm 损失,式 (1) 不再以  $(w_g, h_g)$  作为预测框尺寸的回归目标,而是能够调整预测框的最优尺寸  $(w_p^*, h_p^*)$  以获得更高的 IoU.但是式 (1) 还没有使预测框尺寸达到 IoU 最优.在第 2.2 节里,将给出在式 (1) 的启发下发现的一种可与 IoU 协调的光滑损失函数.

## 2.2 光滑 IoU 损失及预测框中心与尺寸最优关系

沿用  $B_p(w_p, h_p; x_p, y_p)$  和  $B_g(w_g, h_g; x_g, y_g)$  分别表示预测框和真值框,以及  $(x_{\Delta}, y_{\Delta})$  表示预测框的中心偏差.在前文中指出中心偏差  $(x_{\Delta}, y_{\Delta})$  的回归是独立的,不需要考虑预测框尺寸  $(w_p, h_p)$  的情况,而预测框尺寸  $(w_p, h_p)$  在回归过程中需要考虑

到中心偏差  $(x_{\Delta}, y_{\Delta})$  的情况,因此本节主要探讨在预测框尺寸  $(w_p, h_p)$  上的最优关系以及光滑的损失函数.首先构造如下损失函数:

$$\mathcal{L}_{\text{SloU}}(B_p; B_g) = \frac{2|x_{\Delta}| + w_g}{w_p} + \frac{2|y_{\Delta}| + h_g}{h_p} - 2 - \ln \frac{w_g h_g}{w_p h_p} \quad (2)$$

式中,在边界框尺寸定义域  $\mathbf{R}_{++}^2 := \{(x, y) | x > 0, y > 0\}$  内每一个预测框尺寸  $(w_p, h_p) \in \mathbf{R}_{++}^2$  上的偏导数:

$$\frac{\partial \mathcal{L}_{\text{SloU}}}{\partial w_p} = -\frac{2|x_{\Delta}| + w_g}{w_p^2} + \frac{1}{w_p} \quad (3)$$

$$\frac{\partial \mathcal{L}_{\text{SloU}}}{\partial h_p} = -\frac{2|y_{\Delta}| + h_g}{h_p^2} + \frac{1}{h_p} \quad (4)$$

都存在并且在  $\mathbf{R}_{++}^2$  连续,因此式 (2) 在预测框尺寸的全局范围内是连续可微的,并且根据最优性条件令偏导数为 0,可得唯一使式 (2) 最优的预测框尺寸为  $(w_p^*, h_p^*) = \arg \min_{w_p, h_p} \mathcal{L}_{\text{SloU}}(B_p; B_g) = (2|x_{\Delta}| + w_g, 2|y_{\Delta}| + h_g)$ .由此可以看出,使  $\mathcal{L}_{\text{SloU}}$  取最优的是覆盖真值框的最小边界框.

本文定义一种描述预测框中心偏差程度的变量.

**定义 1.** 令  $(x_{\Delta}, y_{\Delta})$  表示预测框中心偏差和  $(w_g, h_g)$  表示真值框尺寸.称  $d_{\mathcal{H}}(x_{\Delta}, y_{\Delta}; w_g, h_g) := |w_g/|x_{\Delta}| - h_g/|y_{\Delta}||$  为预测框中心相对真值框中心的调和归一化偏差.

接下来,阐述 IoU 最优的预测框尺寸与调和归一化偏差  $d_{\mathcal{H}}$  有关.首先讨论式 (2) 构造的光滑损失  $\mathcal{L}_{\text{SloU}}$  在功能上等效于 IoU 损失  $\mathcal{L}_{\text{IoU}}$  的情况.

**命题 1.** 令  $B_p(w_p, h_p; x_p, y_p)$  表示预测框和  $B_g(w_g, h_g; x_g, y_g)$  表示真值框.假设给定了预测框的中心位置  $(x_p, y_p)$ ,且调和归一化偏差  $d_{\mathcal{H}}$  满足  $d_{\mathcal{H}}(x_{\Delta}, y_{\Delta}; w_g, h_g) < 2$ ,那么使式 (2) 在尺寸  $(w_p, h_p) \in \mathbf{R}_{++}^2$  上取得最优的预测框  $B_p^* = (w_p^*, h_p^*; x_p, y_p)$  所取得的  $\text{IoU}(B_p^*; B_g)$  也是最优的,即  $\mathcal{L}_{\text{IoU}}(B_p^*; B_g) = 1 - \text{IoU}(B_p^*; B_g)$  取最小.

**证明.** 已知使式 (2) 达到最优的预测框尺寸为  $(w_p^*, h_p^*) = (2|x_{\Delta}| + w_g, 2|y_{\Delta}| + h_g)$ .根据 IoU 定义  $\text{IoU}(B_p; B_g) = |B_p \cap B_g| / |B_p \cup B_g|$ ,易知此时  $\text{IoU}(B_p^*; B_g) = (w_g h_g) / (w_p^* h_p^*)$ .接下来要证明的是在尺寸上最优化式 (2) 得到的  $B_p^*$  所取得的  $\text{IoU}(B_p^*; B_g)$  是最优的.

记  $C((w_p^*, h_p^*), r) := \{(w, h) \in \mathbf{R}^2; (w - w_p^*)^2 + (h - h_p^*)^2 < r^2\}$  为  $(w_p^*, h_p^*)$  的某个邻域,这样  $(w_p, h_p) \in C((w_p^*, h_p^*), r) \setminus \{(w_p^*, h_p^*)\}$  可表示为  $w_p = w_p^* +$

$t \cos \theta$  和  $h_p = h_p^* + t \sin \theta$ ,  $t \in (0, r)$ ,  $\theta \in (0, 2\pi]$ . 接下来, 只需要证明  $\text{IoU}(B_p; B_g) < \text{IoU}(B_p^*; B_g)$ .

1) 当  $\theta \in (0, \pi/2]$  时, 那么显然有:

$$\text{IoU}(B_p; B_g) = \frac{w_g h_g}{(w_p^* + t \cos \theta)(h_p^* + t \sin \theta)} < \frac{w_g h_g}{w_p^* h_p^*} \quad (5)$$

也就是说,  $\text{IoU}(B_p; B_g) < \text{IoU}(B_p^*; B_g)$ .

2) 当  $\theta \in (\pi/2, \pi]$  时, 有:

$$\text{IoU}(B_p; B_g) = \frac{h_g \left( w_g + \frac{t}{2} \cos \theta \right)}{w_g h_g + w_p h_p - h_g \left( w_g + \frac{t}{2} \cos \theta \right)} \quad (6)$$

比较  $\text{IoU}(B_p; B_g)$  和  $\text{IoU}(B_p^*; B_g)$ , 有:

$$\text{IoU}(B_p; B_g) - \text{IoU}(B_p^*; B_g) = \eta \left( (h_g |x_\Delta| - w_g |y_\Delta| + 2|x_\Delta||y_\Delta|) \cos \theta - w_p^* w_g \sin \theta \right) \quad (7)$$

其中

$$\eta = \frac{h_g t}{\left( w_p h_p - \frac{1}{2} h_g t \cos \theta \right) w_p^* h_p^*} > 0 \quad (8)$$

是一个正数, 本文省略了高阶小量  $O(t^2)$ , 当  $t$  取足够小时, 由于  $|w_g/x_\Delta - h_g/y_\Delta| < 2$ , 显然  $h_g |x_\Delta| - w_g |y_\Delta| + 2|x_\Delta||y_\Delta| > 0$ , 所以  $\text{IoU}(B_p; B_g) < \text{IoU}(B_p^*; B_g)$ .

3) 当  $\theta \in (\pi, 3\pi/2]$  时, 有:

$$\text{IoU}(B_p; B_g) = \frac{w_g \left( h_g + \frac{t}{2} \sin \theta \right)}{w_g h_g + w_p h_p - w_g \left( h_g + \frac{t}{2} \sin \theta \right)} \quad (9)$$

替换  $\theta$  为  $\phi + 2\pi$ ,  $\phi \in (2\pi, \pi]$ , 得到:

$$\text{IoU}(B_p; B_g) = \frac{w_g \left( h_g + \frac{t}{2} \cos \phi \right)}{w_g h_g + w_p h_p - w_g \left( h_g + \frac{t}{2} \cos \phi \right)} \quad (10)$$

与式 (6) 等价, 所以  $\text{IoU}(B_p; B_g) < \text{IoU}(B_p^*; B_g)$ .

4) 当  $\theta \in (3\pi/2, 2\pi]$  时:

$$\text{IoU}(B_p; B_g) = \frac{\left( w_g + \frac{t}{2} \cos \theta \right) \left( h_g + \frac{t}{2} \sin \theta \right)}{w_g h_g + w_p h_p - \left( w_g + \frac{t}{2} \cos \theta \right) \left( h_g + \frac{t}{2} \sin \theta \right)} \quad (11)$$

接着, 比较  $\text{IoU}(B_p; B_g)$  和  $\text{IoU}(B_p^*; B_g)$ , 有:

$$\begin{aligned} \text{IoU}(B_p; B_g) - \text{IoU}(B_p^*; B_g) = & \zeta (h_g (h_g |x_\Delta| - w_g |y_\Delta| + 2|x_\Delta||y_\Delta|) \cos \theta - \\ & w_g (w_g |y_\Delta| - h_g |x_\Delta| + 2|x_\Delta||y_\Delta|) \sin \theta) \quad (12) \end{aligned}$$

其中

$$\zeta = \frac{t}{\left( w_p h_p - \frac{1}{2} w_g t \sin \theta - \frac{1}{2} h_g t \cos \theta \right) w_p^* h_p^*} > 0 \quad (13)$$

是一个正数. 由于  $h_g |x_\Delta| - w_g |y_\Delta| + 2|x_\Delta||y_\Delta| > 0$  以及  $w_g |y_\Delta| - h_g |x_\Delta| + 2|x_\Delta||y_\Delta| > 0$ , 所以  $\text{IoU}(B_p; B_g) < \text{IoU}(B_p^*; B_g)$ .

综上所述, 对于尺寸  $(w_p, h_p)$  最优的预测框  $B_p^*(w_p^*, h_p^*; x_p, y_p) = \arg \min_{w_p, h_p} \mathcal{L}_{\text{SIoU}}(B_p; B_g) = \arg \max_{w_p, h_p} \text{IoU}(B_p; B_g)$  在最小化式 (2) 的同时, 能够最大化 IoU 指标.  $\square$

图 3 给出了一个  $\mathcal{L}_{\text{IoU}} = 1 - \text{IoU}$  和式 (2) 中  $\mathcal{L}_{\text{SIoU}}$  在满足  $d_{\mathcal{H}} < 2$  条件下在相同点取最优的可视化示例, 示例中真值框尺寸为  $(w_g, h_g) = (10, 10)$  以及中心偏差为  $(x_\Delta, y_\Delta) = (2, 2.5)$ , 水平坐标面表示预测框尺寸. 由图 3 可以看出, 两者均在预测框尺寸为  $(w_p^*, h_p^*) = (14, 15)$  处取得最小值. 命题 1 指出如果中心偏差满足  $d_{\mathcal{H}} < 2$ , 则覆盖真值框的最小边界框即是 IoU 最优的. 当中心偏差为  $(0, 0)$  时, 最

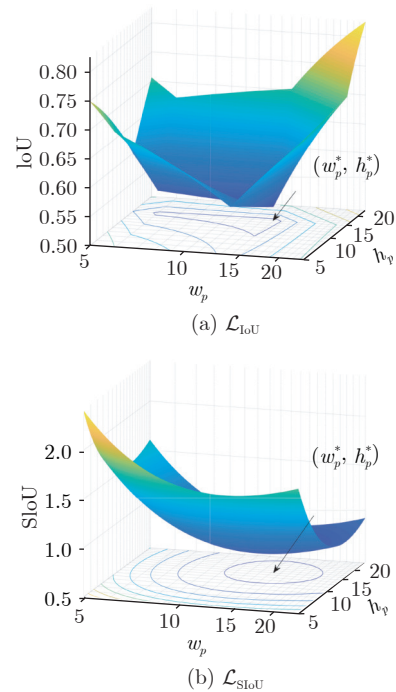


图 3  $\mathcal{L}_{\text{IoU}}$  和  $\mathcal{L}_{\text{SIoU}}$  在对数坐标下的可视化图像示例  
Fig. 3 A visualized example of  $\mathcal{L}_{\text{SIoU}}$  and  $\mathcal{L}_{\text{IoU}}$  viewed in the logarithmic scale of horizontal axis

小覆盖真值框的边界框尺寸与真值框相同, 此时  $\mathcal{L}_{\text{SIoU}}$  的优化目标退化为真值框本身, 与其他边界框回归损失如  $\ell_n$ -norm 或  $\mathcal{L}_{\text{IoU}}$  的目标相容.

由命题 1 可以推出, 调和归一化偏差  $d_{\mathcal{H}}$  满足其他情况时 IoU 最优的预测框尺寸.

**命题 2.** 假设给定预测框的中心  $(x_p, y_p)$  并满足  $d_{\mathcal{H}} = 2$ , 那么 IoU 最优的预测框尺寸  $(w_p^*, h_p^*)$  在区域  $[w_g - 2|x_{\Delta}|, w_g + 2|x_{\Delta}|] \times [h_g + 2|y_{\Delta}|]$  或  $[w_g + 2|x_{\Delta}|] \times [h_g - 2|y_{\Delta}|, h_g + 2|y_{\Delta}|]$  二者之一上.

**命题 3.** 假设给定预测框中心  $(x_p, y_p)$  并满足  $d_{\mathcal{H}} > 2$ , 那么 IoU 最优的预测框尺寸  $(w_p^*, h_p^*)$  为  $(w_g - 2|x_{\Delta}|, h_g + 2|y_{\Delta}|)$  或  $(w_g + 2|x_{\Delta}|, h_g - 2|y_{\Delta}|)$  二者之一.

仿照命题 1 的证明, 可以证明命题 2 和命题 3, 本文不再赘述. 上述 3 个命题揭示了在回归过程中 IoU 最优的边界框各参数之间蕴含的定量关系. 命题 2 指出了  $\mathcal{L}_{\text{IoU}}$  最优的预测框尺寸不唯一的情况. 在边界框中心处于特殊位置 (即  $d_{\mathcal{H}} = 2$ ) 时,  $\mathcal{L}_{\text{IoU}}$  增加了边界框形状或尺寸在回归过程中的不确定性.  $\mathcal{L}_{\text{SIoU}}$  则不存在这个问题, 最优化  $\mathcal{L}_{\text{SIoU}}$  所取得的真值框的最小覆盖框是唯一的.

虽然命题 3 指出了在预测框中心偏差满足  $d_{\mathcal{H}} > 2$  时, 仍以真值框的最小覆盖框作为动态回归目标不是 IoU 最优的, 但是注意到最优化  $\mathcal{L}_{\text{SIoU}}$  仍然可以取得一个良好的预测框. 图 4 显示了一个中心偏差落入  $d_{\mathcal{H}} > 2$  的示例, 实线框代表以  $g_c$  为中心的真值框, 而其余虚线框代表以  $p_c$  为中心的预测框. 在中心距离  $d$  相同的情况下, 图 4(a) 所显示的为依据 IoU 指标最优的预测框, 但是其 GIoU 指标相对较低. 图 4(b) 所显示的为  $\mathcal{L}_{\text{SIoU}}$  最优的边界框, 虽然其 IoU 指标略微低于 IoU 最优的边界框, 但是其 GIoU 指标则显著高于图 4(a). 另外, 值得注意的是, 如果引入额外的先验知识 (例如  $\text{CIoU}^{[6]}$ ) 将预测框限定为保持与真值框相同的宽高比, 其不同尺度的边界框如图 4(c) 所示, 可以看到其所能达到的 IoU 上界低于  $\mathcal{L}_{\text{SIoU}}$  最优的边界框所取得的 IoU 指标. 因此, 即便遵循  $\mathcal{L}_{\text{SIoU}}$  最优所得到的最小覆盖真值的边界框在 IoU 意义下不是最优的, 但是综合 IoU 和 GIoU 指标来看, 依然不失为一个很好的策略.

概括地说, 式 (2) 给出的损失函数  $\mathcal{L}_{\text{SIoU}}$  具有以下特性:

1) 尺度不变性. 与  $\mathcal{L}_{\text{IoU}}$  一样,  $\mathcal{L}_{\text{SIoU}}$  仍然是回归尺度不变的损失函数. 尺度不变是指在损失相同的情况下预测框与真值框之间的 IoU 不会随着边界框尺度的变化而变化. 相对于尺度变化的损失函数

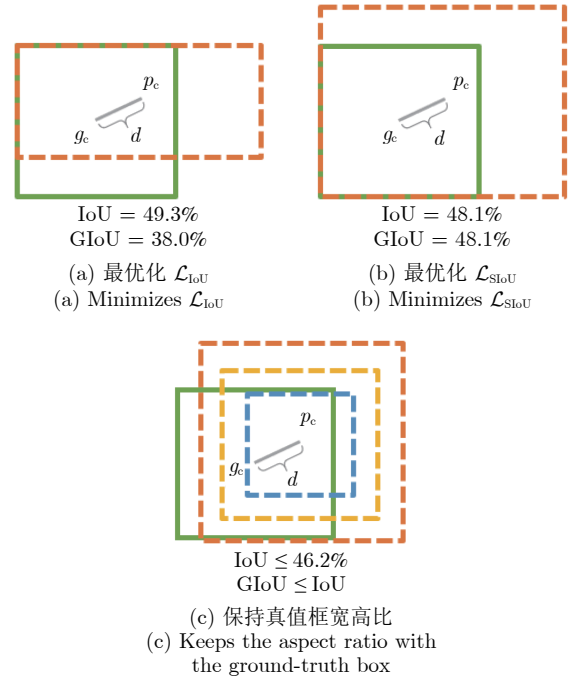


图 4 当  $d_{\mathcal{H}} > 2$  时最优化  $\mathcal{L}_{\text{SIoU}}$  和  $\mathcal{L}_{\text{IoU}}$  的边界框示例

Fig. 4 Illustration of predicted box that minimizes  $\mathcal{L}_{\text{SIoU}}$  and  $\mathcal{L}_{\text{IoU}}$  if  $d_{\mathcal{H}} > 2$

例如  $\ell_n$ -norm 损失, 尺度不变的损失函数可以减轻目标尺寸的多样性带来的不利影响.

2) 正定性. 当且仅当预测框与真值框完全重合时, 即 IoU 指标达到最大值 1 时,  $\mathcal{L}_{\text{SIoU}} = 0$  达到最小值. 由此,  $\mathcal{L}_{\text{SIoU}}$  可以视为一种散度函数反映出预测框与真值框的相近程度, 或者更准确地说,  $\mathcal{L}_{\text{SIoU}}$  反映出以预测框代替真值框而产生的损失程度.

3) 光滑性与极值唯一性. 当预测框与真值框不相交时, 有  $\mathcal{L}_{\text{IoU}} = 1$ , 此时  $\nabla \mathcal{L}_{\text{IoU}} = \mathbf{0}$ , 无法通过梯度下降更新预测框参数; 并且  $\mathcal{L}_{\text{IoU}}$  取最优时  $\nabla \mathcal{L}_{\text{IoU}}$  不存在, 导致回归的结果不稳定. 而  $\mathcal{L}_{\text{SIoU}}$  在全局上偏导数存在且连续, 预测框参数可以通过梯度下降更新, 且更容易回归到极值处, 当且仅当  $\mathcal{L}_{\text{SIoU}}$  取最优时, 有  $\nabla \mathcal{L}_{\text{SIoU}} = \mathbf{0}$ ,  $\mathcal{L}_{\text{SIoU}}$  达到极值.

### 2.3 预测框中心偏差的正则项

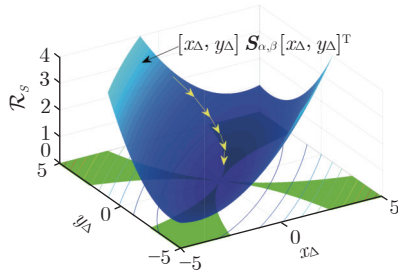
为了使预测框的中心位置在回归过程中尽可能满足条件  $d_{\mathcal{H}} < 2$ , 提出一种针对  $\mathcal{L}_{\text{SIoU}}$  的正则项:

$$\mathcal{R}_S = [x_{\Delta}, y_{\Delta}] \mathbf{S}_{\alpha, \beta} [x_{\Delta}, y_{\Delta}]^T \quad (14)$$

其中

$$\mathbf{S}_{\alpha, \beta} := \begin{bmatrix} \frac{\beta + \cos \alpha}{\sin \alpha} & 1 \\ 1 & \frac{\beta - \cos \alpha}{\sin \alpha} \end{bmatrix} \quad (15)$$

式中,  $\alpha = 2 \arctan(w_g/h_g)$  是由真值框尺寸所确定的参量, 而  $\beta$  是需要满足  $\beta > 1$  的参量. 图 5 给出了一个由正方形真值框确定的正则项  $\mathcal{R}_S$  的图像示例, 阴影区域满足  $d_H < 2$ , 箭头代表某一梯度轨迹. 由于一般的  $\ell_2$  正则项  $\| [x_\Delta, y_\Delta] \|_2^2 = x_\Delta^2 + y_\Delta^2$ , 其等值线是同心圆, 梯度总是指向  $(0, 0)$ , 这样中心偏差  $(x_\Delta, y_\Delta)$  位于平面内任何方向上的机会都是均等的, 并不适用于 SIOU 损失的特点. 由图 5 可以看出, 不同于一般的  $\ell_2$  正则项, 本文针对 SIOU 损失设计的正则项  $[x_\Delta, y_\Delta] \mathbf{S}_{\alpha, \beta} [x_\Delta, y_\Delta]^T$ , 其关联的非对角正定的二次型矩阵  $\mathbf{S}_{\alpha, \beta}$  使  $[x_\Delta, y_\Delta] \mathbf{S}_{\alpha, \beta} [x_\Delta, y_\Delta]^T$  的等值线为椭圆形并且长轴恰位于直线  $d_H = 0$ . 所以正则项  $[x_\Delta, y_\Delta] \mathbf{S}_{\alpha, \beta} [x_\Delta, y_\Delta]^T$  在等值线上的梯度指向是不同的, 具有方向偏好. 如果以梯度下降法更新  $(x_\Delta, y_\Delta)$ , 则  $(x_\Delta, y_\Delta)$  的轨迹可以向着  $d_H < 2$  区域靠拢如图 5 所示, 增大落入到区域  $d_H < 2$  的机会, 同时也可以增加  $\mathcal{L}_{\text{SIOU}}$  的凸性. 注意到正则项  $[x_\Delta, y_\Delta] \mathbf{S}_{\alpha, \beta} [x_\Delta, y_\Delta]^T$  是一个仅和中心偏差  $(x_\Delta, y_\Delta)$  有关的函数, 而与预测框尺寸  $(w_p, h_p)$  无关, 所以正则项的加入并不会使  $\mathcal{L}_{\text{SIOU}}$  违反最佳尺寸与中心偏差的关系.

图 5 正则项  $\mathcal{R}_S$  的图像示例Fig. 5 Illustration of regularization  $\mathcal{R}_S$ 

加入中心偏差的正则项后, 提出的光滑 IoU 损失函数  $\mathcal{L}_{\text{SIOU}}$  可如下表示为:

$$\mathcal{L}_{\text{SIOU}}(B_p; B_g) = \frac{w_g + 2|x_\Delta|}{w_p} + \frac{h_g + 2|y_\Delta|}{h_p} - \ln \frac{w_g h_g}{w_p h_p} - 2 + \gamma [x_\Delta, y_\Delta] \mathbf{S}_{\alpha, \beta} [x_\Delta, y_\Delta]^T \quad (16)$$

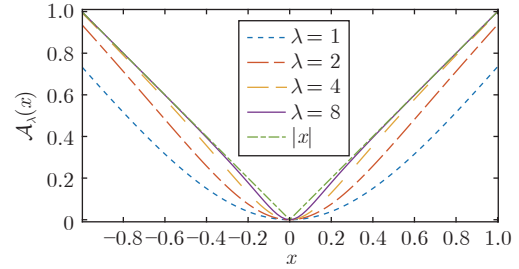
式中,  $\gamma$  为正则项的系数.

## 2.4 预测框中心偏差的光滑代理

注意到  $\mathcal{L}_{\text{SIOU}}$  里含有关于中心偏差的  $\ell_1$  函数  $|x_\Delta|$  和  $|y_\Delta|$ , 其在 0 处也是不可微的. 针对这个问题, 本文不妨构造一个近似函数以替换  $|x_\Delta|$  和  $|y_\Delta|$ . 考虑到对于任意  $x_\Delta \in \mathbf{R}$ , 当  $\lambda \rightarrow \infty$  有下式成立:

$$\mathcal{A}_\lambda(x_\Delta) := \frac{1}{\lambda} \ln (e^{\lambda x_\Delta} + e^{-\lambda x_\Delta} - 1) \xrightarrow{U} |x_\Delta| \quad (17)$$

式中,  $\xrightarrow{U}$  代表一致收敛. 易知  $\mathcal{A}_\lambda(x_\Delta)$  是光滑的, 图 6 给出了取不同  $\lambda$  值的光滑代理函数  $\mathcal{A}_\lambda(x_\Delta)$  的图像. 这样  $\mathcal{A}_\lambda(x_\Delta)$  可以用来作为  $|x_\Delta|$  的光滑代理.  $|x_\Delta|$  和  $|y_\Delta|$  在回归过程中可以分别用  $\mathcal{A}_\lambda(x_\Delta)$  和  $\mathcal{A}_\lambda(y_\Delta)$  代替以保证  $\mathcal{L}_{\text{SIOU}}$  对边界框中心位置参数是光滑的.

图 6 不同参数  $\lambda$  下  $|x|$  的光滑代理函数  $\mathcal{A}_\lambda(x)$ Fig. 6 Plot of smooth surrogate function  $\mathcal{A}_\lambda(x)$  for  $|x|$  with different  $\lambda$  controlling its shape

## 2.5 光滑 IoU 损失的算法

在应用光滑 IoU 损失训练边界框回归时需要注意其中两点: 1) 容易验证当  $(w_p, h_p)$  趋于  $(0, 0)$  时,  $\mathcal{L}_{\text{SIOU}}$  损失则趋于无穷, 这样在训练初期可能因为预测的尺寸过小而出现梯度爆炸的情况. 为了避免训练过程中的梯度爆炸, 对  $\mathcal{L}_{\text{SIOU}}$  作了梯度截断处理, 通过取:

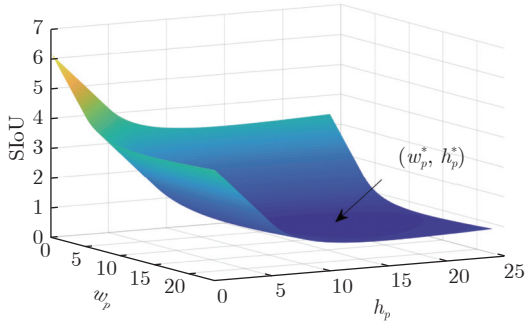
$$F(w_p; w_g) = \left( \frac{4|x_\Delta|}{w_g} + 1 \right) \left( 1 + \left| 1 - \frac{2w_p}{w_g} \right| \right) - \ln 2 \quad (18)$$

$$F(h_p; h_g) = \left( \frac{4|y_\Delta|}{h_g} + 1 \right) \left( 1 + \left| 1 - \frac{2h_p}{h_g} \right| \right) - \ln 2 \quad (19)$$

与  $\mathcal{L}_{\text{SIOU}}$  最小的操作:

$$\min (\mathcal{L}_{\text{SIOU}}(B_p; B_g), F(w_p; w_g) + F(h_p; h_g)) \quad (20)$$

使预测框尺寸在  $w_p \leq w_g/2$ ,  $h_p \leq h_g/2$  时梯度不变, 限制到可控范围内, 同时不影响  $\mathcal{L}_{\text{SIOU}}$  的可微性. 图 7 给出了以图 3 中的示例通过梯度截断后的  $\mathcal{L}_{\text{SIOU}}$  损失图像. 2) 为了避免光滑代理函数  $\mathcal{A}_\lambda$  中的指数函数可能引发机器浮点数溢出, 不妨设置一个区间半径  $r_\lambda > 0$ , 当  $-r_\lambda \leq x \leq r_\lambda$ , 取  $\mathcal{A}_\lambda(x)$ , 否则取  $|x|$ . 选取适当的区间半径  $r_\lambda$  可以在机器所能表示的精度范围内保持连续性. 应用光滑 IoU 损失训练边界框回归如算法 1 所示. 光滑 IoU 损失可以很容易代替 IoU 损失函数应用在深度目标跟踪网络中训练边界框的回归. 在下一节里将组织相关实验以验证提出的光滑 IoU 损失的有效性.

图 7 梯度截断后的  $\mathcal{L}_{\text{SIoU}}$  可视化示例Fig.7 A visualized example of  $\mathcal{L}_{\text{SIoU}}$  with truncated gradient**算法 1.** 应用光滑 IoU 损失的边界框回归

**输入.** 边界框回归分支预测的阳性位置到 4 条边线的距离向量  $(t_p, l_p, b_p, r_p)$ , 有监督的 4 边真值距离向量  $(t_g, l_g, b_g, r_g)$ .

**输出.** 光滑 IoU 损失  $\mathcal{L}_{\text{SIoU}}$ .

- 1) 计算真值框尺寸:  $(w_g, h_g) = (l_g + r_g, t_g + b_g)$ ;
- 2) 计算预测框尺寸:  $(w_p, h_p) = (l_p + r_p, t_p + b_p)$ ;
- 3) 计算预测框相对真值框的中心偏差:

$$(x_\Delta, y_\Delta) = \frac{1}{2}(r_p - l_p - r_g + l_g, t_p - b_p - t_g + b_g)$$

- 4) 取  $|x_\Delta|$  和  $|y_\Delta|$  的代理函数  $\mathcal{A}_\lambda(x_\Delta)$  和  $\mathcal{A}_\lambda(y_\Delta)$ :

$$\mathcal{A}_\lambda(x_\Delta) = \begin{cases} \frac{1}{\lambda} \ln(e^{\lambda x_\Delta} + e^{-\lambda x_\Delta} - 1), & |x_\Delta| < r_\lambda \\ |x_\Delta|, & \text{其他} \end{cases}$$

$$\mathcal{A}_\lambda(y_\Delta) = \begin{cases} \frac{1}{\lambda} \ln(e^{\lambda y_\Delta} + e^{-\lambda y_\Delta} - 1), & |y_\Delta| < r_\lambda \\ |y_\Delta|, & \text{其他} \end{cases}$$

- 5) 梯度截断处理:

$$T(w_p) = \min\left(\frac{w_g + 2\mathcal{A}_\lambda(x_\Delta)}{w_p} - \ln \frac{w_g}{w_p} - 1, F(w_p)\right)$$

$$T(h_p) = \min\left(\frac{h_g + 2\mathcal{A}_\lambda(y_\Delta)}{h_p} - \ln \frac{h_g}{h_p} - 1, F(h_p)\right)$$

- 6) 计算预测框中心偏差的正则项:

$$\mathcal{R}_S = (x_\Delta, y_\Delta) \mathbf{S}_{\alpha, \beta}(x_\Delta, y_\Delta)^T, \beta = 2, \alpha = 2 \arctan \frac{w_g}{h_g}$$

- 7) 计算 SIoU 损失:  $\mathcal{L}_{\text{SIoU}} = T(w_p) + T(h_p) + \gamma \mathcal{R}_S$ .

### 3 实验和评估

本节将提出的光滑 IoU 损失合并到具有代表性的基于锚点无关的目标跟踪模型 SiamFC++<sup>[20]</sup>、SiamBAN<sup>[21]</sup> 和 SiamCAR<sup>[22]</sup> 中来评估其有效性. 其原本的 IoU 损失  $\mathcal{L}_{\text{IoU}}$  训练的结果作为基线, 用  $\mathcal{L}_{\text{SIoU}}$  替换原本的  $\mathcal{L}_{\text{IoU}}$  训练作为对比. 实验中选择 GoogleNet<sup>[25]</sup> 作为 SiamFC++<sup>[20]</sup> 的孪生骨干网络结果.

而 SiamBAN<sup>[21]</sup> 和 SiamCAR<sup>[22]</sup> 则采用 ResNet-50<sup>[26]</sup> 的后三个残差块级联的方式提取特征, 并遵循 SiamFC++、SiamBAN 和 SiamCAR 的训练过程, 采用论文里报告的默认参数和每个基准上的迭代次数. 实验中  $\lambda$  设为 2,  $r_\lambda$  设为 20. 正则项系数设置为  $\gamma = 1/\sqrt{w_g h_g}$ . 采用 ILSVRC-VID/DET<sup>[27]</sup>、COCO<sup>[28]</sup>、YoutubeBB<sup>[29]</sup>、LaSOT<sup>[30]</sup>、TrackingNet<sup>[31]</sup> 和 GOT-10k<sup>[32]</sup> 作为基础训练集. 然后在主流的目标跟踪测评基准平台 LaSOT<sup>[30]</sup>、TrackingNet<sup>[31]</sup>、GOT-10k<sup>[32]</sup>、OTB2015<sup>[33]</sup> 和 VOT2018<sup>[34]</sup> 上, 对提出的边界框回归损失模型进行评估对比. 另外, 仅从跟踪结果上很难讨论边界框回归的过程是如何进行. 因此本节设置了一组采样分析实验, 采样的数据综合考虑了距离、尺度以及宽高比等边界框之间的关系, 涵盖多种回归情况, 研究光滑 IoU 损失相比当前基于 IoU 的损失的优越性. 实验环境配备了 128 GB 内存, Intel Xeon E5-2650 2.3 GHz CPU 处理器, Nvidia GTX 1080Ti GPU 显卡, 采用深度学习框架 PyTorch 实现.

#### 3.1 采样实验

遵照文献 [6-7], 本文设置了一组实验来验证提出的边界框回归损失的高效性. 为了尽可能探索不同的边界框在不同中心距离、尺度以及宽高比等方面的回归情况, 从均值协方差不同的 2 个二元正态分布  $N(25, 10^2)$ 、 $N(50, 30^2)$ , 分别采样  $N = 5000$  个坐标作为初始预测框集合  $\{B_p^j\}_{j=1}^{N \times S}$  的中心点代表初始预测框距离真值框较近和较远的情况, 其中  $S = 3 \times 5$  代表初始预测框覆盖 3 种不同尺寸, 分别为 0.8、1.0 和 1.2, 并且每种尺度分别设置 5 种形状的初始宽高比, 分别是 1:4、1:2、1:1、2:1 和 4:1. 真值框集合  $\{B_g^i\}_{i=1}^5$  同样设定了这 5 种宽高比, 其中心点均位于 (0, 0), 如图 8 所示. 全部  $5000 \times 5 \times 3 = 75000$  个初始预测框要回归到每一种真值框, 总共得到  $5000 \times 5 \times 3 \times 5 = 375000$  个回归情况. 依据梯度下降的方向更新预测框的变量  $B_p^j \leftarrow B_p^j - \gamma \nabla_{B_p^j} \mathcal{L}_{\text{SIoU}}(B_p^j; B_g^i)$ . 回归过程的效率依据 IoU 指标  $\sum_{i,j} \text{IoU}(B_p^j, B_g^i)$ , GIoU 指标  $\sum_{i,j} \text{GIoU}(B_p^j, B_g^i)$  和基于  $l_2$  的回归误差  $\sum_{i,j} \|B_p^j - B_g^i\|_2^2$  进行评估.

由图 9 可以看出, 本文 SIoU 损失在初期的迭代中收敛得更快速, 同时可以取得更高的 IoU 指标、GIoU 指标, 以及更低的  $l_2$  回归误差. 图 10 展示了不同回归方法在相同迭代过程的截面图, 图 10 中真值框参数为 (5, 20; 10, 40), 而预测框初始参数为 (40, 52; 28, 14), 左、中和右分别代表采用  $\mathcal{L}_{\text{GIoU}}$ 、 $\mathcal{L}_{\text{CIoU}}$  和本文提出的  $\mathcal{L}_{\text{SIoU}}$  在第 10 次、第 40 次和第

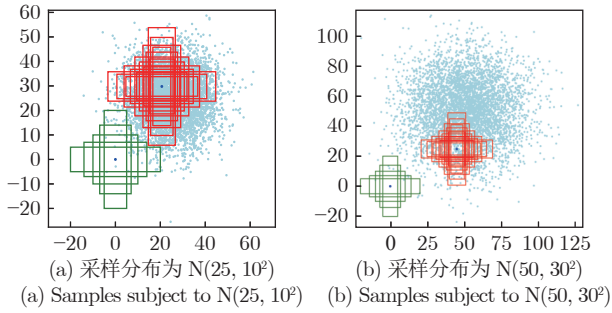


图 8 从两种分布中采样近距离和远距离的初始预测框位置

Fig.8 Sample the initial predicted boxes subject to normal distribution with short and long mean-variance

120 次迭代得到的预测框. 可以看出, 相较于其他基于 IoU 的损失函数,  $\mathcal{L}_{\text{SIoU}}$  与其他的损失函数采取的策略不同, 比如  $\mathcal{L}_{\text{CIoU}}$  是先较快地将预测框尺寸回归到真值框的尺寸, 但这样就与真值框的重叠较少; 而  $\mathcal{L}_{\text{GIoU}}$  虽然可以快速增大预测框尺寸使其包含真值框, 但中心距离收敛得较慢, 所以它们的 IoU 指标增加得缓慢. 而  $\mathcal{L}_{\text{SIoU}}$  在回归过程中是 IoU 友好的, 能够更快促进边界框发生交叠. 也就是说, 即便在回归初期预测框位置离真值框较远时,  $\mathcal{L}_{\text{SIoU}}$  依然可以快速将预测框的尺寸增大到与边界框产生交叠, 然后通过收缩边界框之间的中心距离以及尺寸来降低  $\ell_2$  回归误差和提高 IoU 指标.

### 3.2 对比实验

本节通过 5 个主流的基准测试来评估提出的光滑 IoU 损失函数, 用于目标边界框回归的性能.

#### 3.2.1 LaSOT

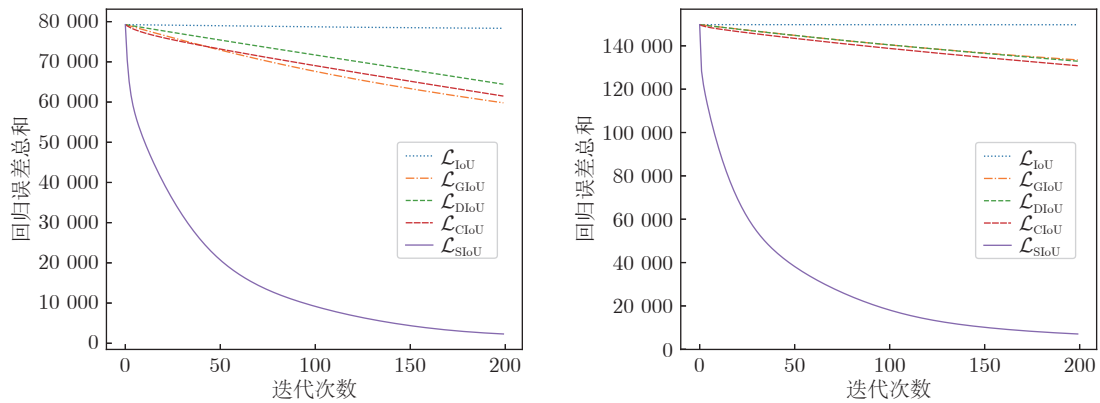
LaSOT<sup>[30]</sup> 是一个高质量的大规模单目标跟踪基准, 数据集包含 1400 个视频, 涵盖了视觉跟踪里 14 种典型的挑战, 例如遮挡、运动模糊、尺度变化等, 划分为 70 个常见类别, 每个类别提供 20 个视频, 平均视频长度超过 2500 帧, 总共超过 352 万个人工标注的帧. LaSOT 基准的协议将其中 1120 个视频作为训练集, 280 个视频作为测试集, 每个类别包含相同数量的视频. 大规模的训练集使得跟踪器不容易出现过拟合, 从而达到了测试跟踪器真实性能的目的. 遵照 LaSOT 基准的协议, 跟踪器需在 LaSOT 训练集上训练, 并在 LaSOT 测试子集上评估, 常用的评估指标为一次性通过 (One-pass evaluation, OPE) 的标准化精确率图、精确率图和成功率图, 其中精确率图刻画了预测边界框与标定边界框的中心位置的像素距离在阈值范围内的图像

帧数所占的比率关系, 精确率以中心位置误差小于 20 像素的比率对跟踪器进行排名; 成功率图刻画了预测的边界框与标定的边界框的重叠率 (即 IoU) 超过阈值的图像帧数所占的比率关系, 然后依据曲线下区域面积对跟踪器进行排名. 而标准化精确率引自 TrackingNet<sup>[31]</sup>, 是为了消除精确率对图像分辨率和边界框尺寸过于敏感, 可表示为  $P_{norm} = \sqrt{(x_{\Delta}/w_g)^2 + (y_{\Delta}/h_g)^2}$ . 表 1 给出了 SiamFC++ 模型以  $\mathcal{L}_{\text{SIoU}}$  作为边界框回归损失训练在 LaSOT 上得到的测试结果. 可以看出,  $\mathcal{L}_{\text{SIoU}}$  相对于原有的  $\mathcal{L}_{\text{IoU}}$  提高了 SiamFC++ 模型性能的成功率、精确率、标准化精确率, 分别相对提高 3.60%、5.05% 和 3.24%. 从 LaSOT 数据集中选择了 5 个代表不同类型的视频, 从中抽取部分帧来显示跟踪的效果, 如图 11 所示. 图中虚线框标出了以  $\mathcal{L}_{\text{SIoU}}$  训练的测试结果, 点线框标出了原始的以  $\mathcal{L}_{\text{IoU}}$  训练的测试结果, 实线框为真值框. 可以看出,  $\mathcal{L}_{\text{SIoU}}$  比  $\mathcal{L}_{\text{IoU}}$  得到的预测框更靠近真值框. 为了验证本文光滑 IoU 损失在其他深度目标跟踪器上也具有良好的鲁棒性和适用性, 表 2 和表 3 分别报告了对 SiamBAN<sup>[21]</sup> 和 SiamCAR<sup>[22]</sup> 模型采用  $\mathcal{L}_{\text{SIoU}}$  替换原有的 IoU 损失训练的实验对比结果. 鉴于不同模型其网络结构的不同, 虽然未能超过 SiamFC++ 的表现, 但  $\mathcal{L}_{\text{SIoU}}$  相对于原有的  $\mathcal{L}_{\text{IoU}}$  提高了 SiamBAN 和 SiamCAR 模型的性能, 其中提升最显著的成功率分别相对提高 5.64% 和 5.04%.

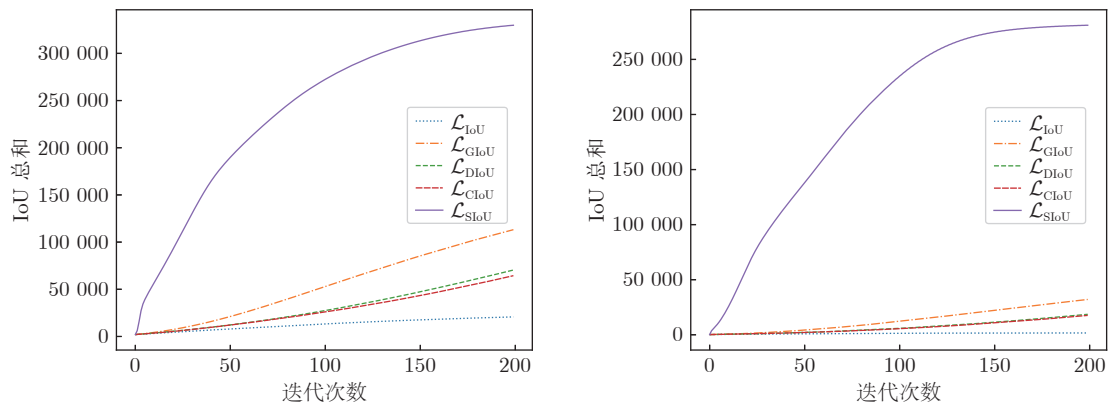
为了表现采用本文 SIoU 损失训练 SiamFC++、SiamBAN 和 SiamCAR 后横向比较的性能, 选取了其他 9 种先进的跟踪方法进行对比, 分别为 SiamBAN<sup>[21]</sup>、SiamCAR<sup>[22]</sup>、SiamRPN++<sup>[9]</sup>、SiamMask<sup>[35]</sup>、GlobalTrack<sup>[36]</sup>、C-RPN<sup>[37]</sup>、Ocean<sup>[23]</sup>、ATOM<sup>[38]</sup> 和 DiMP<sup>[39]</sup>. 其中, Ocean<sup>[23]</sup> 模型选择具有更好性能的在线更新的 Ocean-online 版本. 用 SiamFC++ (SIoU) 代表以  $\mathcal{L}_{\text{SIoU}}$  训练边界框回归分支的版本以示区分. 在 LaSOT 上的成功率和精确率的对比如图 12 和表 4 所示, 可以看出, SIoU 损失方法使 SiamFC++ 模型超越了先进的 Ocean 和 DiMP, 实现了最好的性能. 与 Ocean-online 相比, SiamFC++ (SIoU) 在 3 个指标上的得分分别相对提高了 1.5%、1.6% 和 1.8%. 与 DiMP 相比, SiamFC++ (SIoU) 在成功率上同样表现出 1.5% 的优势, 而在精度上表现出 1.9% 的优势, 验证了  $\mathcal{L}_{\text{SIoU}}$  可以更好地在复杂场景中回归不同对象边界框的能力.

#### 3.2.2 GOT-10k

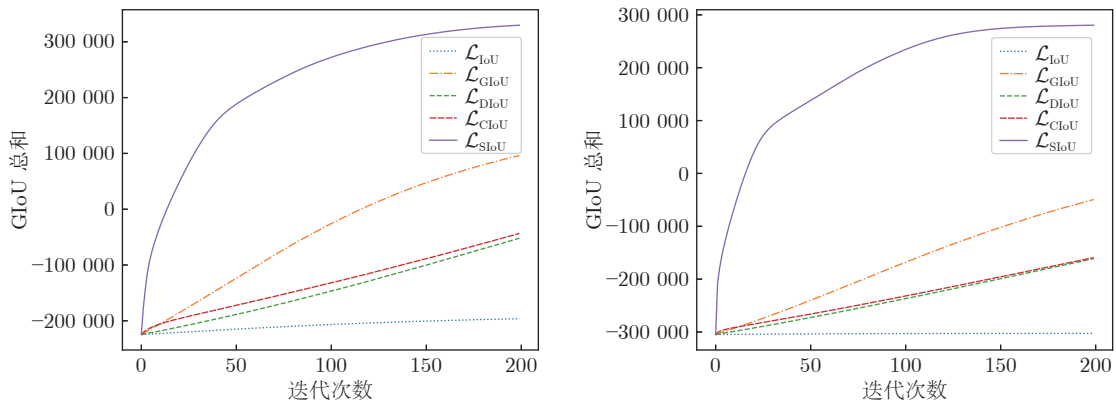
GOT-10k<sup>[32]</sup> 一个由中国科学院发布的基于 Wo-



(a) 不同边界框损失在近距离(左)和远距离(右)的回归误差曲线  
(a) Plots of short-range (left) and long-range (right) regression error on different regression losses



(b) 不同边界框损失在近距离(左)和远距离(右)的IoU曲线  
(b) Plots of short-range (left) and long-range (right) IoU metric on different regression losses



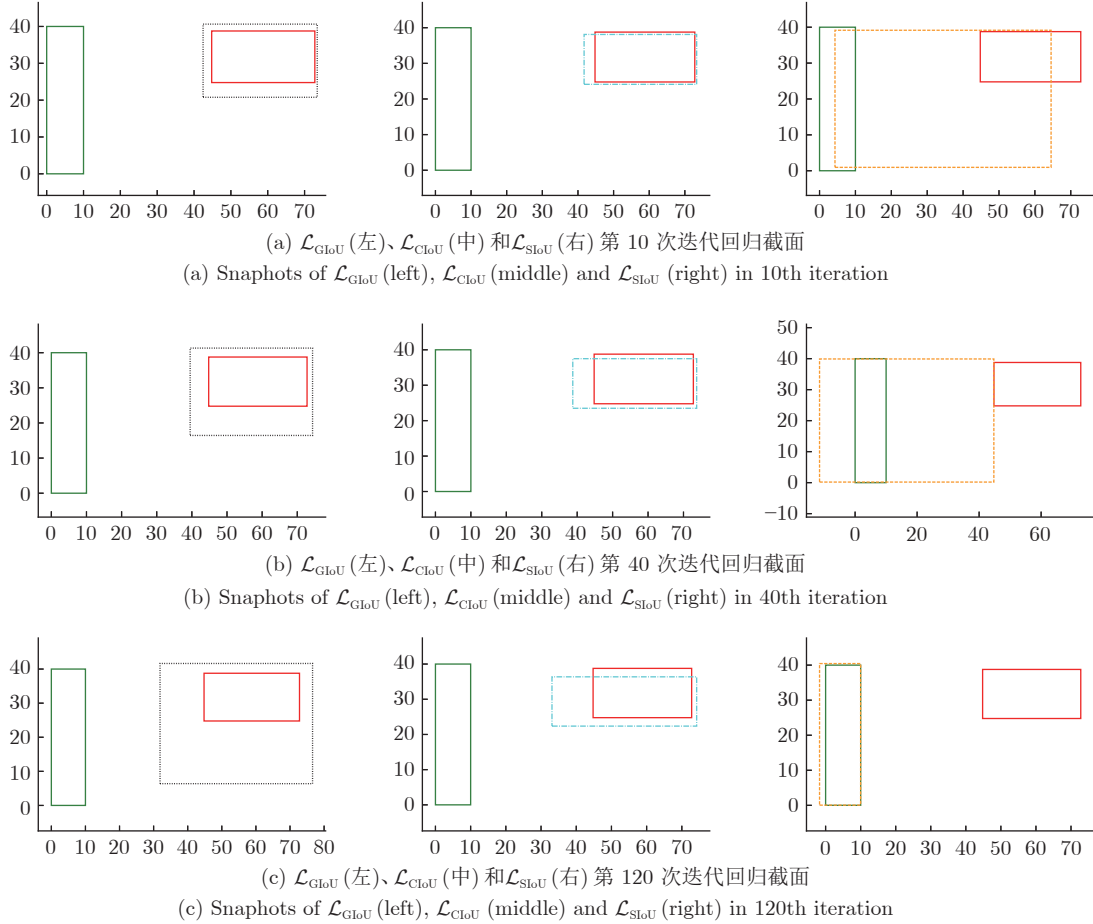
(c) 不同边界框损失在近距离(左)和远距离(右)的GIoU曲线  
(c) Plots of short-range (left) and long-range (right) GIoU metric on different regression losses

图9 各种边界框回归损失比较

Fig.9 Comparison among the convergence performance of different bounding box regression losses

rdNet 的大型目标跟踪数据集, 总共超过 10000 段视频, 细分了 563 类户外常见的移动物体, 范围涵盖了动物、交通工具、人物、被动运动目标以及特定部位目标 5 大类别, 标注的边界框数量超过 150 万。除了类别广泛, 规模宏大, 该数据集还具有训练数据统一和单样本学习等特点。依照 GOT-10k 的协

议, 所有模型都用相同的训练数据, 来保障所有模型之间的公平对比。并且为了使训练出的模型能有更强的泛化能力, 基准测试集与训练集之间不存在交集。测试集包含 180 段视频, 分属于 84 个目标类别, 该测试基准评价的指标有平均重叠率 (Average overlap, AO) 和成功率 (Success rate, SR), 数

图 10 不同迭代次数的  $\mathcal{L}_{\text{GIoU}}$ 、 $\mathcal{L}_{\text{CIoU}}$  和  $\mathcal{L}_{\text{SIoU}}$  的回归示例Fig.10 Illustration of predicted boxes via  $\mathcal{L}_{\text{GIoU}}$ ,  $\mathcal{L}_{\text{CIoU}}$  and  $\mathcal{L}_{\text{SIoU}}$  regressing in different iterations表 1 在基准 LaSOT 上, 分别以  $\mathcal{L}_{\text{IoU}}$  (原本的) 和  $\mathcal{L}_{\text{SIoU}}$  训练的模型 SiamFC++ 的测试结果 (%)Table 1 Comparison between the performance of SiamFC++ trained using  $\mathcal{L}_{\text{IoU}}$  (original),  $\mathcal{L}_{\text{SIoU}}$  on the test set of LaSOT (%)

评价指标	成功率	精确度	标准化精确度
$\mathcal{L}_{\text{IoU}}$	55.6	55.5	64.8
$\mathcal{L}_{\text{SIoU}}$	57.6	58.3	66.9
相对增益	3.60	5.05	3.24

值越大说明方法性能越高. 表 5 展示了采用  $\mathcal{L}_{\text{SIoU}}$  训练的 SiamFC++ 模型在服务器上评估的结果, 虽然在  $\text{SR}_{0.75}$  指标上性能略低于原始结果 2.29%, 但在  $\text{SR}_{0.50}$  指标上增益高达 7.48% 以及在 AO 指标上增益 3.69%, 同样实现了一定程度的性能改进. 表 6 给出了经过  $\mathcal{L}_{\text{SIoU}}$  训练的 SiamCAR 模型在服务器上的评估结果, 3 个指标上均有不同程度的提升, 除了在  $\text{SR}_{0.50}$  指标上性能提升最高达 6.29%, 在 AO 和  $\text{SR}_{0.75}$  指标上相对基线结果分别提升了

3.69% 和 5.22%. 表 7 总结了与 7 种当前先进方法 MDNet<sup>[40]</sup>、SPM<sup>[19]</sup>、ATOM<sup>[38]</sup>、SiamCAR<sup>[22]</sup>、SiamRPN++<sup>[9]</sup>、Ocean-online<sup>[23]</sup>、D3S<sup>[41]</sup> 和 DiMP-50<sup>[39]</sup> 的对比, SiamFC++ (SIoU) 和 SiamCAR (SIoU) 代表以  $\mathcal{L}_{\text{SIoU}}$  训练边界框回归分支的版本. 可以看出, 采用  $\mathcal{L}_{\text{SIoU}}$  训练的 SiamFC++ 在标准化精确度和成功率方面都表现出了优势.  $\mathcal{L}_{\text{SIoU}}$  使 SiamFC++ 的成功率超过了最先进的 DiMP-50<sup>[39]</sup> 和 Ocean-online<sup>[23]</sup> 达 1.8%, 标准化精确度超过了 2.2%.

### 3.2.3 TrackingNet

为了进一步评估本文方法, 在更具挑战性的数据集 TrackingNet<sup>[31]</sup> 上进行了实验. TrackingNet 包含了 30 132 个视频, 平均每个视频 471.4 帧, 以及覆盖了 27 个类别用于单目标跟踪器的训练, 是目前目标跟踪任务里的体量最大的数据集. 与 GOT-10k 类似, TrackingNet 的测试集独立于训练集, 并在官方评估服务器上测试, 该基准测试提供了 511 个视频, 视频平均帧数与类别属性分布与训练集相似. 与基准 LaSOT 相同, 评估服务器基于跟

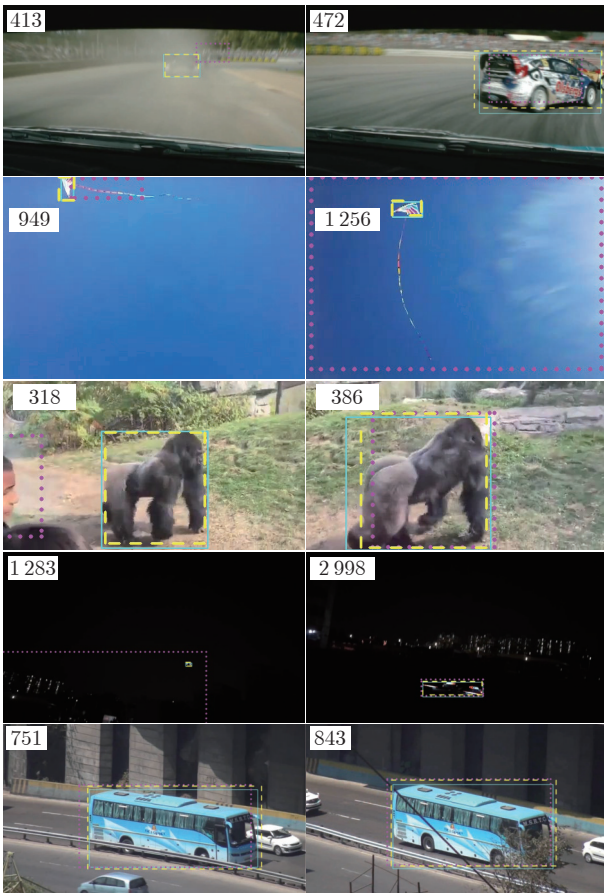


图 11 在 LaSOT 测试集上, 分别以  $\mathcal{L}_{IoU}$  (点线框标出) 和  $\mathcal{L}_{Siou}$  (虚线框标出) 训练的模型 SiamFC++ 的可视化结果示例 (实线框为真值标签)

Fig. 11 Visualized tracking results of SiamFC++ trained using  $\mathcal{L}_{IoU}$  (marked in dotted box) and  $\mathcal{L}_{Siou}$  (marked in dashed box) on LaSOT (solid box denotes groundtruth)

表 2 在基准 LaSOT 上, 分别以  $\mathcal{L}_{IoU}$  (原本的) 和  $\mathcal{L}_{Siou}$  训练的模型 SiamBAN 的测试对比 (%)

Table 2 Comparison between the performance of SiamBAN trained using  $\mathcal{L}_{IoU}$  (original),  $\mathcal{L}_{Siou}$  on the test set of LaSOT (%)

评价指标	成功率	精确度	标准化精确度
$\mathcal{L}_{IoU}$	51.4	52.1	59.8
$\mathcal{L}_{Siou}$	54.3	53.9	63.3
相对增益	5.64	3.45	4.85

踪结果计算成功率、精确度和标准化精确度三个评估指标. 表 8 给出了 SiamFC++ 模型以  $\mathcal{L}_{Siou}$  作为边界框回归损失训练在 TrackingNet 上得到的测试结果. 可以看出,  $\mathcal{L}_{Siou}$  相对于原有的  $\mathcal{L}_{IoU}$  提高了 SiamFC++ 模型的性能, 成功率、精确度以及标准化精确度分别相对提高 1.06%、2.27% 和 2.37%.

表 3 在基准 LaSOT 上, 分别以  $\mathcal{L}_{IoU}$  (原本的) 和  $\mathcal{L}_{Siou}$  训练模型 SiamCAR 的测试对比 (%)

Table 3 Comparison between the performance of SiamCAR trained using  $\mathcal{L}_{IoU}$  (original),  $\mathcal{L}_{Siou}$  on the test set of LaSOT (%)

评价指标	成功率	精确率	标准化精确率
$\mathcal{L}_{IoU}$	51.6	52.4	61.0
$\mathcal{L}_{Siou}$	54.9	54.8	63.1
相对增益	6.39	4.58	3.44

而表 9 给出了与 7 种当前先进的跟踪器, 即 MD-Net<sup>[40]</sup>、ATOM<sup>[38]</sup>、DaSiamRPN<sup>[10]</sup>、Siam-RPN++<sup>[9]</sup>、UpdateNet<sup>[42]</sup>、SPM<sup>[19]</sup>、DiMP<sup>[39]</sup> 在 TrackingNet 上的结果对比, SiamFC++ (Siou) 代表以  $\mathcal{L}_{Siou}$  训练边界框回归分支的版本. 可以看出, 采用  $\mathcal{L}_{Siou}$  训练的 SiamFC++ 在精确度和成功率方面均表现最佳. 如表 9 所示,  $\mathcal{L}_{Siou}$  使 SiamFC++ 的成功率超过了最先进的 DiMP<sup>[39]</sup> 模型 1.8%, 标准化精确率超过了 2.2%. 在如此大规模的数据集上的实验结果表明了以 Siou 损失训练边界框回归具有良好的泛化能力.

### 3.2.4 OTB2015

除了上述大规模数据的基准测试, 本文也在小规模的数据集 OTB2015<sup>[33]</sup> 上进行了实验. OTB2015 包含了 100 个视频, 涵盖了视觉跟踪里 11 种典型的挑战. 与 LaSOT 类似, 该基准测试常用的评估指标为一次性通过的精确率和成功率. 表 10 和表 11 分别给出了以  $\mathcal{L}_{Siou}$  作为边界框回归损失训练的 SiamFC++ 模型和 SiamBAN 模型在 OTB2015 上得到的测试结果. 可以看出, 虽然  $\mathcal{L}_{Siou}$  相对于原有的  $\mathcal{L}_{IoU}$  提高了 SiamFC++ 模型和 SiamBAN 模型的性能, 但提升幅度有限, 成功率相对提高了分别为 0.74% 和 0.43%, 而精确率相对提高了分别为 0.34% 和 0.55%. 可能的原因是小规模测试集对网络参数以及超参数更敏感, 具有偶然性和特殊性, 大规模的测试样本更能得到一般性的结果.

### 3.2.5 VOT2018

数据集 VOT2018<sup>[34]</sup> 共包含 60 个视频, 虽然视频数量较少并与 VOT2018 之前版本发布的数据集相同, 但是对所有视频重新标定了由分割掩码外接得到的更加精确的边界框, 也就是说这种边界框不再是坐标轴对齐的, 给跟踪器带来了新的挑战. VOT2018 里重要的 3 个评价指标: 准确率 (Accuracy, A)、鲁棒性 (Robustness, R) 和平均重叠率期望 (Expected average overlap, EAO). 准确率用来评价跟踪器的准确度, 通过  $n$  次重复测试得到跟踪器在单个视频帧序列下 IoU 的平均值, 即  $A =$

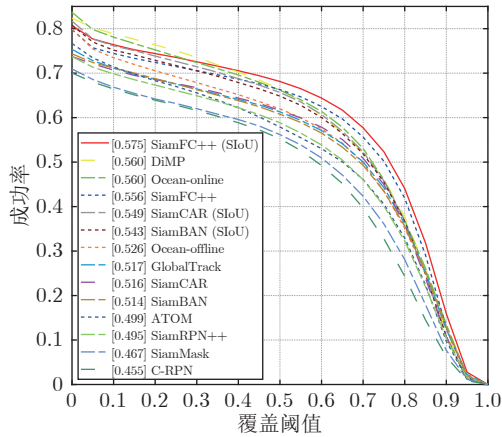
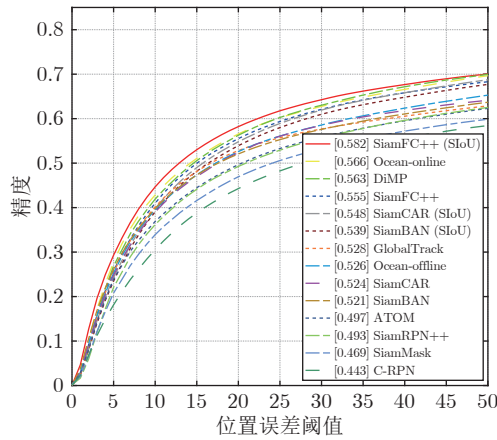
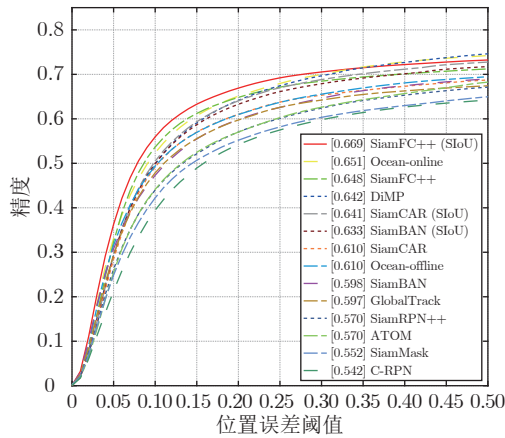
(a) 成功率图  
(a) Success plot(b) 精确率图  
(b) Precision plot(c) 标准化精确率图  
(c) Normalized precision plot

图 12 在 LaSOT 上评估成功率、精确率和标准化精确率结果

Fig.12 Success plot with area under the curve, precision plot and normalized precision plot on LaSOT

$1/n \sum_{i=1}^n \text{IoU}(i)$ , 该指标数值越大, 准确度越高. 鲁棒性用来评价跟踪器的稳定性, 通过  $n$  次重复测试

表 4 在基准 LaSOT 上, 与先进方法的性能评估对比  
Table 4 Performance evaluation for state-of-the-art algorithms on LaSOT

方法	成功率	精确率	标准化精确率
SiamBAN	51.4	52.1	59.8
ATOM	51.5	50.5	57.6
SiamCAR	51.6	52.4	61.0
SiamRPN++	49.6	49.1	56.9
Ocean-online	56.0	56.6	65.1
SiamFC++	55.6	55.5	64.8
DiMP	56.8	56.4	64.3
SiamBAN (SIoU)	54.3	53.9	63.3
SiamCAR (SIoU)	54.2	53.7	63.1
SiamFC++ (SIoU)	57.6	58.3	66.9

表 5 在 GOT-10k 上, 分别以  $\mathcal{L}_{\text{IoU}}$  (原本的) 和  $\mathcal{L}_{\text{SIoU}}$  训练的模型 SiamFC++ 测试对比 (%)Table 5 Comparison between the performance of SiamFC++ trained using  $\mathcal{L}_{\text{IoU}}$  (original),  $\mathcal{L}_{\text{SIoU}}$  on the test set of GOT-10k (%)

评价指标	AO	SR <sub>0.50</sub>	SR <sub>0.75</sub>
$\mathcal{L}_{\text{IoU}}$	59.5	69.5	47.9
$\mathcal{L}_{\text{SIoU}}$	61.7	74.7	46.8
相对增益	3.69	7.48	-2.29

表 6 在 GOT-10k 上, 分别以  $\mathcal{L}_{\text{IoU}}$  (原本的) 和  $\mathcal{L}_{\text{SIoU}}$  训练的模型 SiamCAR 测试结果 (%)Table 6 Comparison between the performance of SiamCAR trained using  $\mathcal{L}_{\text{IoU}}$  (original),  $\mathcal{L}_{\text{SIoU}}$  on the test set of GOT-10k (%)

评价指标	AO	SR <sub>0.50</sub>	SR <sub>0.75</sub>
$\mathcal{L}_{\text{IoU}}$	58.1	68.3	44.1
$\mathcal{L}_{\text{SIoU}}$	60.2	72.6	46.4
相对增益	3.61	6.29	5.22

得到跟踪器在单个视频帧序列上跟踪失败的次数  $F$  的平均值, 即  $R = 1/n \sum_{i=1}^n F(i)$ , 重叠率为 0 即为跟踪失败, 该指标数值越小, 稳定性越高. VOT-2018 相较于其他测试基准具有的一个特色机制是会在跟踪器跟踪失败时重启, 即失败发生时的 5 帧后重新初始化, 所以平均重叠率期望是取跟踪器在非重新初始化的  $N_l$  个长度为  $l$  的视频帧序列上平均重叠率的期望值, 即  $\text{EAO} = \text{E}_l[1/N_l \sum_{i=1}^{N_l} \text{IoU}_l(i)]$  是 VOT2018 评估跟踪算法精度的重要指标, 数值越大, 精度越高. 表 12 给出了 SiamFC++

表 7 在基准 GOT-10k 上, 与先进方法的性能评估对比 (%)

Table 7 Performance evaluation for state-of-the-art algorithms on GOT-10k (%)

方法	AO	SR <sub>0.50</sub>
MDNet	29.9	30.3
SPM	51.3	59.3
ATOM	55.6	63.4
SiamCAR	56.9	67.0
SiamRPN++	51.7	61.8
Ocean-online	61.1	72.1
D3S	59.7	67.6
SiamFC++	59.5	69.5
DiMP-50	61.1	71.2
SiamCAR (SIoU)	60.2	72.6
SiamFC++ (SIoU)	61.7	74.7

表 8 在 TrackingNet 上, 分别以  $\mathcal{L}_{IoU}$  (原本的) 和  $\mathcal{L}_{SIoU}$  训练的模型 SiamFC++ 的测试结果 (%)

Table 8 Comparison between the performance of SiamFC++ trained using  $\mathcal{L}_{IoU}$  (original),  $\mathcal{L}_{SIoU}$  on the test of TrackingNet (%)

评价指标	精确率	标准化精确率	成功率
$\mathcal{L}_{IoU}$	70.5	80.0	75.4
$\mathcal{L}_{SIoU}$	72.1	81.9	76.2
相对增益	2.27	2.37	1.06

表 9 在基准 TrackingNet 上, 与先进方法的性能评估对比 (%)

Table 9 Performance evaluation for state-of-the-art algorithms on TrackingNet (%)

方法	成功率	标准化精确率
MDNet	60.6	70.5
ATOM	70.3	77.1
DaSiamRPN	63.8	73.3
SiamRPN++	73.3	80.0
UpdateNet	67.7	75.2
SPM	71.2	77.8
SiamFC++	75.4	80.0
DiMP	74.0	80.1
SiamFC++ (SIoU)	76.2	81.9

模型以  $\mathcal{L}_{SIoU}$  作为边界框回归损失训练在 VOT2018 上的测试结果. 由表 12 可以看出, 在准确率和 EAO 指标上有所下降. 造成这种现象可能的原因是, VOT2018 里的 IoU 计算涉及到预测框与旋转

表 10 在 OTB2015 上, 分别以  $\mathcal{L}_{IoU}$  (原本的) 和  $\mathcal{L}_{SIoU}$  训练的模型 SiamFC++ 的测试结果 (%)

Table 10 Comparison between the performance of SiamFC++ trained using  $\mathcal{L}_{IoU}$  (original),  $\mathcal{L}_{SIoU}$  on the test of OTB2015 (%)

评价指标	成功率	标准化精确率
$\mathcal{L}_{IoU}$	68.2	89.5
$\mathcal{L}_{SIoU}$	68.7	89.8
相对增益	0.74	0.34

表 11 在 OTB2015 上, 分别以  $\mathcal{L}_{IoU}$  (原本的) 和  $\mathcal{L}_{SIoU}$  训练的模型 SiamBAN 测试结果 (%)

Table 11 Comparison between the performance of SiamBAN trained using  $\mathcal{L}_{IoU}$  (original),  $\mathcal{L}_{SIoU}$  on on the test of OTB2015 (%)

评价指标	成功率	标准化精确率
$\mathcal{L}_{IoU}$	69.6	91.0
$\mathcal{L}_{SIoU}$	69.9	91.5
相对增益	0.43	0.55

表 12 在 VOT2018 上, 分别以  $\mathcal{L}_{IoU}$  (原本的) 和  $\mathcal{L}_{SIoU}$  训练的模型 SiamFC++ 测试结果 (%)

Table 12 Comparison between the performance of SiamFC++ trained using  $\mathcal{L}_{IoU}$  (original),  $\mathcal{L}_{SIoU}$  on on the test of VOT2018 (%)

评价指标	准确率	鲁棒性	EAO
$\mathcal{L}_{IoU}$	0.586	0.201	0.427
$\mathcal{L}_{SIoU}$	0.582	0.196	0.400

的标注框之间的交叠, 而非传统意义下两个坐标轴对齐的矩形框之间的 IoU, 而此时并不能证明提出的光滑 IoU 损失所遵从的策略仍然可以最优. 所以为了应对这种评估指标, 还有待对边界框回归函数做进一步研究和拓展.

### 3.3 消融实验

为了证明提出的光滑 IoU 损失与其他以 IoU 为基准的损失如  $\mathcal{G}IoU$ <sup>[5]</sup> 和  $\mathcal{D}IoU$ <sup>[6]</sup> 相比具有优势, 本文在基准 LaSOT 和 GOT-10k 上对 SiamFC++、SiamBAN 和 SiamCAR 模型采取不同边界框回归损失函数 (即本文提出的  $\mathcal{L}_{SIoU}$ 、 $\mathcal{L}_{\mathcal{G}IoU}$ <sup>[5]</sup>、 $\mathcal{L}_{\mathcal{D}IoU}$ <sup>[6]</sup> 以及原本的  $\mathcal{L}_{IoU}$  损失) 作为对比实验. 表 13 和表 14 记录了基于 3 种模型的不同边界框回归损失在测试集上超过不同 IoU 阈值的图像帧数所占比率. 最小阈值取值为 0.5, 并以步幅 0.05 逐次累加的方式设置更高的阈值, 可以看出提出的光滑 IoU 损失可以改善边界框回归的效果, 虽然在高 IoU 阈值下其 SIoU 训练的模型测试结果所占比率不一定高于其

表 13 在基准 LaSOT 上, 与其他基于 IoU 损失训练得到的满足不同 IoU 阈值的测试集图像帧数占比的对比结果 (%)  
Table 13 Comparison results with other IoU-based loss for the ratio of frames exceeding different IoU thresholds on the test set of LaSOT (%)

IoU 阈值	$\geq 0.95$	$\geq 0.90$	$\geq 0.85$	$\geq 0.80$	$\geq 0.75$	$\geq 0.70$	$\geq 0.65$	$\geq 0.60$	$\geq 0.55$	$\geq 0.50$
SiamFC++ (SIoU)	<b>2.75</b>	15.93	<b>31.83</b>	<b>44.05</b>	<b>52.33</b>	<b>57.71</b>	<b>61.71</b>	<b>64.41</b>	<b>66.52</b>	<b>68.14</b>
SiamFC++ (DIOU)	1.60	13.19	29.72	42.84	51.48	57.09	61.28	64.17	66.31	67.95
SiamFC++ (GIOU)	2.45	<b>16.18</b>	31.10	42.26	50.39	55.81	59.56	62.37	64.58	66.34
SiamFC++	1.52	12.73	29.10	41.90	50.20	55.63	59.72	62.51	64.68	66.31
SiamBAN (SIoU)	1.18	10.79	<b>24.86</b>	<b>36.64</b>	<b>45.50</b>	<b>51.87</b>	<b>56.45</b>	<b>60.03</b>	<b>62.77</b>	<b>64.84</b>
SiamBAN (GIOU)	1.49	11.77	24.71	35.15	44.77	50.79	54.93	57.76	60.70	63.98
SiamBAN	<b>1.98</b>	<b>12.89</b>	25.40	35.57	43.46	49.29	53.38	56.53	58.92	60.78
SiamCAR (SIoU)	1.20	10.80	24.81	<b>36.74</b>	<b>45.75</b>	<b>52.29</b>	<b>56.97</b>	<b>60.71</b>	<b>63.53</b>	<b>65.66</b>
SiamCAR (DIOU)	1.20	<b>10.91</b>	<b>25.10</b>	36.62	45.04	51.47	56.18	59.89	62.70	64.83
SiamCAR	<b>1.27</b>	10.62	23.90	35.98	44.93	50.87	55.08	57.86	59.94	61.61

表 14 在基准 GOT-10k 上, 与其他基于 IoU 损失训练得到的满足不同 IoU 阈值的测试集图像帧数占比的对比结果 (%)  
Table 14 Comparison results with other IoU-based loss for the ratio of frames exceeding different IoU thresholds on the test set of GOT-10k (%)

IoU 阈值	$\geq 0.95$	$\geq 0.90$	$\geq 0.85$	$\geq 0.80$	$\geq 0.75$	$\geq 0.70$	$\geq 0.65$	$\geq 0.60$	$\geq 0.55$	$\geq 0.50$
SiamFC++ (SIoU)	0.94	<b>8.18</b>	<b>22.01</b>	<b>35.76</b>	<b>46.83</b>	<b>55.71</b>	<b>62.16</b>	<b>67.40</b>	<b>71.39</b>	<b>74.68</b>
SiamFC++ (DIOU)	0.72	7.56	21.10	35.11	46.86	55.45	61.68	66.82	70.74	73.86
SiamFC++ (GIOU)	<b>0.97</b>	7.20	21.80	34.19	45.85	54.50	59.24	63.48	66.02	69.49
SiamCAR (SIoU)	0.94	<b>8.58</b>	<b>22.20</b>	<b>35.83</b>	<b>46.46</b>	<b>54.74</b>	<b>60.66</b>	<b>65.49</b>	<b>69.30</b>	<b>72.62</b>
SiamCAR (GIOU)	<b>1.13</b>	6.72	19.23	34.78	45.37	53.73	59.98	64.96	68.95	71.96
SiamCAR (DIOU)	0.92	6.19	18.85	32.54	43.73	52.51	58.86	64.04	68.09	71.33
SiamCAR	0.81	8.02	20.76	33.88	44.07	51.87	57.21	61.35	64.99	68.31

他 IoU 为基准的损失, 但是这一部分的比率普遍很小, SIoU 损失在中高 IoU 阈值下与其他基于 IoU 的损失相比优势明显, 对整体指标提升的贡献更大.

最后, 为了探讨中心偏差的正则项  $\mathcal{R}_S$  和光滑代理函数  $\mathcal{A}_\lambda$  所带来性能上的影响, 本文在 GOT-10k 上对其进行了消融实验. 表 15 报告了不同消融的结果, 其中  $\mathcal{L}_{\text{SIoU}}(\text{w/o } \mathcal{A}\mathcal{R})$  代表不具有正则项  $\mathcal{R}_S$  和光滑代理函数  $\mathcal{A}_\lambda$  的损失,  $\mathcal{L}_{\text{SIoU}}(\text{w/ } \mathcal{R})$  代表配备了正则项  $\mathcal{R}_S$  而不采用光滑代理函数  $\mathcal{A}_\lambda$  的损失, 而  $\mathcal{L}_{\text{SIoU}}(\text{w/ } \mathcal{A}_\lambda)$  代表采用了光滑代理函数  $\mathcal{A}_\lambda$  而不配备正则项  $\mathcal{R}_S$  的损失. 在本文  $\lambda$  取 1、2、4 和 8 四个值以观察不同  $\lambda$  的影响. 由表 15 和图 13 可以看出, 正则项和代理函数提高了边界框回归损失的性能. 其中对中心偏差的正则项可以较好弥补  $\mathcal{L}_{\text{SIoU}}$  在  $d_{\mathcal{H}} > 2$  时与 IoU 不匹配带来的差异. 同时也注意到, 相较于正则项的加入, 将中心偏差  $(x_\Delta, y_\Delta)$  的损失从  $\ell_1$ -norm 替换为光滑代理函数  $\mathcal{A}_2$ 、 $\mathcal{A}_4$  和  $\mathcal{A}_8$  所带来的性能增益有限, 其中可能的原因是不同于边界框尺寸的回归目标是动态的, 边界框中心位置的回归目标是静态的, 总是指向真值框的中心, 也

就是  $(x_\Delta, y_\Delta)$  的优化目标总是  $(0, 0)$ , 但实际上中心偏差  $(x_\Delta, y_\Delta)$  很难回归到  $(0, 0)$ , 因此对边界框中心位置的回归作光滑处理带来的增益较小. 至于  $\lambda$  取值为 1 时, 结果却逊于  $|x|$ , 可能是因为  $\mathcal{A}_\lambda(x)$  与  $|x|$  的误差较大如图 6 所示, 所以  $\lambda$  取值适中即可, 既不必取值太小使  $\mathcal{A}_\lambda(x)$  偏差  $|x|$  较大, 也不必取值太大使  $\mathcal{A}_\lambda(x)$  在原点处过于“尖锐”而失去光滑的意义, 本文不妨取值为 2.

## 4 结束语

本文给出并证明了在回归过程中最优边界框参数之间满足的定量关系, 提出了一种新的用于训练边界框回归的损失, 即光滑 IoU 损失. 该光滑 IoU 损失不以 IoU 损失作为基本损失, 从优化散度的角度构造了全局光滑且极值唯一的损失函数, 提出的光滑 IoU 损失蕴含边界框各参数之间特定的最优关系, 并将边界框参数作为一个整体进行回归, 其唯一极值可使 IoU 达到最优. 该损失函数确保了在全局上可梯度下降更新参数, 使得边界框更容易回归到极值处, 从而规避了 IoU 损失的固有缺陷. 在

表 15 在 GOT-10k 上, 对  $\mathcal{L}_{\text{IoU}}$  的正则项和代理函数的消融实验 (%)

Table 15 Ablation studies about the regularization and surrogate function on GOT-10k (%)

评价指标	AO	SR <sub>0.50</sub>	SR <sub>0.75</sub>
$\mathcal{L}_{\text{IoU}}$ (w/o $\mathcal{AR}$ )	59.9	72.1	46.3
$\mathcal{L}_{\text{IoU}}$ (w/ $\mathcal{AR}$ )	61.7	74.7	46.8
相对增益	3.01	3.61	1.08
$\mathcal{L}_{\text{IoU}}$ (w/ $\mathcal{R}$ )	61.4	74.5	46.7
相对增益	2.51	3.33	0.86
$\mathcal{L}_{\text{IoU}}$ (w/ $\mathcal{A}_2$ )	60.3	72.6	46.5
相对增益	0.67	0.69	0.43
$\mathcal{L}_{\text{IoU}}$ (w/ $\mathcal{A}_4$ )	60.6	73.4	46.3
相对增益	1.17	1.80	0
$\mathcal{L}_{\text{IoU}}$ (w/ $\mathcal{A}_8$ )	60.4	73.4	46.0
相对增益	0.83	1.80	-0.65
$\mathcal{L}_{\text{IoU}}$ (w/ $\mathcal{A}_1$ )	58.9	71.3	43.7
相对增益	-1.17	-1.11	-5.62

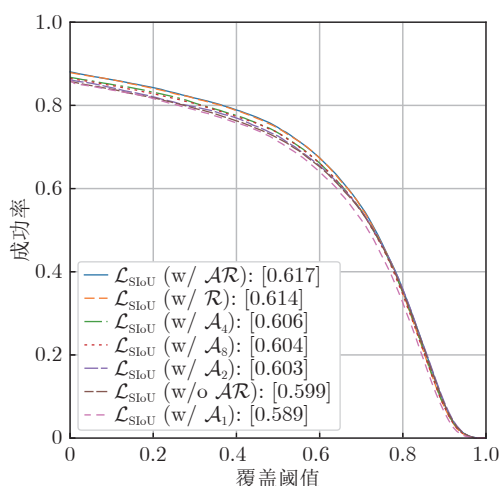


图 13 在 GOT-10k 上的成功率图

Fig. 13 Success plot on GOT-10k

采样数据上进行的大量实验表明, 光滑 IoU 损失和现有基于 IoU 的损失方法相比, 收敛速度更快, 带来了显著的改进. 光滑 IoU 损失可以很容易地集成到当前基于 IoU 损失的视觉任务模型中, 本文将其应用在具有代表性的无锚框目标跟踪模型 SiamFC++、SiamBAN 和 SiamCAR 上, 在 LaSOT、GOT-10k、TrackingNet 和 OTB2015 等主流测试基准上所取得的结果验证了光滑 IoU 损失可以帮助提高边界框回归模块的性能.

## References

1 Meng Lu, Yang Xu. A survey of object tracking algorithms. *Acta Automatica Sinica*, 2019, **45**(7): 1244–1260

(孟球, 杨旭. 目标跟踪算法综述. *自动化学报*, 2019, **45**(7): 1244–1260)

2 Jiang Hong-Yi, Wang Yong-Juan, Kang Jin-Yu. A survey of object detection models and its optimization methods. *Acta Automatica Sinica*, 2021, **47**(6): 1232–1255  
(蒋弘毅, 王永娟, 康锦煜. 目标检测模型及其优化方法综述. *自动化学报*, 2021, **47**(6): 1232–1255)

3 Girshick R B. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1440–1448

4 Yu J, Jiang Y, Wang Z, Cao Z, Huang T S. Unitbox: An advanced object detection network. In: Proceedings of the ACM Conference on Multimedia Conference. Amsterdam, Netherland: 2016. 516–520

5 Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I D, Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 658–666

6 Zheng Z, Wang P, Li J, Ye R, Ren D. Distance-IOU loss: Faster and better learning for bounding box regression. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020. 12993–13000

7 Zhang Y, Ren W, Zhang Z, Jia Z, Wang L, Tan T. Focal and efficient IOU loss for accurate bounding box regression [Online], available: <https://arxiv.org/abs/2101.08158>, August 11, 2021

8 Li B, Yan J, Wu W, Zhu Z, Hu X. High performance visual tracking with siamese region proposal network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: 2018. 8971–8980

9 Li B, Wu W, Wang Q, Wu W, Yan J, Hu W. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 4282–4291

10 Zhu Z, Wang Q, Li B, Wu W, Yan J, Hu W. Distractor-aware siamese networks for visual object tracking. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: 2018. 103–119

11 He Y, Zhu C, Wang J, Savvides M, Zhang X. Bounding box regression with uncertainty for accurate object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 2888–2897

12 Law H, Deng J. Cornernet: Detecting objects as paired keypoints. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: 2018. 765–781

13 Gidaris S, Komodakis N. Locnet: Improving localization accuracy for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: 2016. 789–798

14 Zhou X, Koltun V, Krähenbühl P. Tracking objects as points. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: 2020. 474–490

15 Lin T Y, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **42**(2): 318–327

16 Held D, Thrun S, Savarese S. Learning to track at 100 FPS with deep regression networks. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands: 2016. 749–765

17 Bertinetto L, Valmadre J, Henriques J F, Vedaldi A, Torr H S

- P. Fully-convolutional siamese networks for object tracking. In: Proceedings of the European Conference on Computer Vision Workshops. Amsterdam, Netherlands: 2016. 850–865
- 18 Jiang B, Luo R, Mao J, Xiao T, Jiang Y. Acquisition of localization confidence for accurate object detection. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: 2018. 816–832
- 19 Wang G, Luo C, Xiong Z, Zeng W. SPM-tracker: Series-parallel matching for real-time visual object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 3643–3652
- 20 Xu Y, Wang Z, Li Z, Ye Y, Yu G. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020. 12549–12556
- 21 Chen Z, Zhong B, Li G, Zhang S, Ji R. Siamese box adaptive network for visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 6667–6676
- 22 Guo D, Wang J, Cui Y, Wang Z, Chen S. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 6268–6276
- 23 Zhang Z, Peng H, Fu J, Li B, Hu W. Ocean: Object-aware anchor-free tracking. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: 2020. 771–787
- 24 Tan Jian-Hao, Zheng Ying-Shuai, Wang Yao-Nan, Ma Xiao-Ping. AFST: Anchor-free fully convolutional siamese tracker with searching center point. *Acta Automatica Sinica*, 2021, **47**(4): 801–812  
(谭建豪, 郑英帅, 王耀南, 马小萍. 基于中心点搜索的无锚框全卷积孪生跟踪器. *自动化学报*, 2021, **47**(4): 801–812)
- 25 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: 2015. 1–9
- 26 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: 2016. 770–778
- 27 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, **115**(3): 211–252
- 28 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: 2014. 740–755
- 29 Real E, Shlens J, Mazzocchi S, Pan X, Vanhoucke V. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: 2017. 7464–7473
- 30 Fan H, Lin L, Yang F, Chu P, Deng G, Yu S, et al. Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 5374–5383
- 31 Müller M, Bibi A, Giancola S, Al-Subaihi S, Ghanem B. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: 2018. 310–327
- 32 Huang L, Zhao X, Huang K. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(5): 1562–1577
- 33 Wu Y, Lim J, Yang M. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**(9): 1834–1848
- 34 Kristan M, He Z. The sixth visual object tracking VOT2018 challenge results. In: Proceedings of the European Conference on Computer Vision Workshops. Munich, Germany: 2018. 3–53
- 35 Wang Q, Zhang L, Bertinetto L, Hu W, Torr P H S. Fast online object tracking and segmentation: A unifying approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 1328–1338
- 36 Huang L, Zhao X, Huang K. Globaltrack: A simple and strong baseline for long-term tracking. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020. 11037–11044
- 37 Fan H, Ling H. Siamese cascaded region proposal networks for real-time visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 7952–7961
- 38 Danelljan M, Bhat G, Khan F S, Felsberg M. ATOM: Accurate tracking by overlap maximization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 4660–4669
- 39 Bhat G, Danelljan M, Gool L V, Timofte R. Learning discriminative model prediction for tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 6181–6190
- 40 Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 4293–4302
- 41 Lukezic A, Matas J, Kristan M. D3S: A discriminative single shot segmentation tracker. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 7131–7140
- 42 Zhang L, Gonzalez-Garcia A, Weijervan De J, Danelljan M, Khan F S. Learning the model update for siamese trackers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 4009–4018



李 功 哈尔滨工业大学模式识别与智能系统研究中心博士研究生. 分别于 2015 年和 2018 年获得哈尔滨工业大学学士和硕士学位. 主要研究方向为计算机视觉中的目标跟踪, 模式识别. E-mail: ligong101@126.com

(LI Gong Ph.D. candidate at the Pattern Recognition and Intelligent System Research Center, Harbin Institute of Technology. He received his bachelor and master degrees from Harbin Institute of Technology in 2015 and 2018, respectively. His research interest covers target tracking in computer vision and pattern recognition.)



**赵 巍** 哈尔滨工业大学副教授. 主要研究方向为模式识别, 机器学习和计算机视觉. 本文通信作者.

E-mail: zhaowei@hit.edu.cn

**(ZHAO Wei** Associate professor at Harbin Institute of Technology. Her research interest covers pattern recognition, machine learning, and computer vision. Corresponding author of this paper.)



**刘 鹏** 哈尔滨工业大学教授. 2007 年获得哈尔滨工业大学博士学位. 主要研究方向为图像处理, 视频分析, 模式识别和大规模集成电路设计.

E-mail: pengliu@hit.edu.cn

**(LIU Peng** Professor at Harbin Institute of Technology. He received

his Ph.D. degree from Harbin Institute of Technology in 2007. His research interest covers image processing, video analysis, pattern recognition, and design of large scale integrated circuits.)



**唐降龙** 哈尔滨工业大学教授. 1995 年获得哈尔滨工业大学博士学位. 主要研究方向为模式识别, 图像处理和机器学习. E-mail: tangxl@hit.edu.cn

**(TANG Xiang-Long** Professor at Harbin Institute of Technology. He received his Ph.D. degree from Harbin Institute of Technology in 1995. His research interest covers pattern recognition, image processing, and machine learning.)

bin Institute of Technology in 1995. His research interest covers pattern recognition, image processing, and machine learning.)