

基于一致性加权的目标检测标注融合算法

刘洋^{1,2}, 李国亮^{1,2}

(1. 农业农村部智慧养殖技术重点实验室 农业智能技术教育部工程研究中心;
2. 华中农业大学信息学院, 湖北 武汉 430070)

摘要: 随着目标检测技术的发展与应用,模型对高质量标注数据的需求激增。然而,传统数据标注存在成本高昂、准确性低等问题,严重制约了模型的训练与发展。为此,提出基于一致性加权的目标检测标注融合算法CWF算法,为高效融合目标检测多标注提供了新方法。首先,构建了一个4 000张牛只图像的数据集,并由多人进行目标检测标注;其次,引入一致性加权机制CWF算法提升标注融合性能;最后,通过融合不同数量的标注结果,测试不同标注融合算法的性能差异与最佳融合规模。实验表明,所提方法相较于交集、WBF、EM等融合算法优势明显,Macro-F1值为91.49%(最高),推理速度相较于精度接近的EM融合算法快4.54倍,兼顾了高精度与推理速度。

关键词: 标注融合;目标检测;牛只图像数据集;弱监督

DOI:10.11907/rjtk.261023

扫描二维码阅读全文:

中图分类号:TP391

文献标识码:A

文章编号:1672-7800(XXXX)0XX-0001-09



Consistency Weight-Based Annotation Fusion Algorithm for Object Detection

LIU Yang^{1,2}, LI Guoliang^{1,2}

(1. Key Laboratory of Intelligent Breeding Technology of the Ministry of Agriculture and Rural Affairs, Agricultural Intelligent Technology Engineering Research Center of the Ministry of Education; 2. College of Informatics, Huazhong Agricultural University, Wuhan 430070, China)

Abstract: With the advancement and application of object detection technology, there is a surge in the demand for high-quality labeled data. However, traditional data annotation suffers from issues such as high costs and low accuracy, severely constraining model training and development. To address this, a consistency-weighted object detection annotation fusion algorithm (CWF) is proposed, offering a novel approach for efficiently integrating multiple annotations in object detection. First, a dataset of 4 000 cattle images was constructed and annotated by multiple annotators. Second, the consistency-weighted mechanism (CWF algorithm) was introduced to enhance annotation fusion performance. Finally, by fusing varying numbers of annotations, the performance differences and optimal fusion scale of different annotation fusion algorithms were tested. Experiments demonstrate that the proposed method significantly outperforms intersection-based, WBF, and EM fusion algorithms, achieving a Macro-F1 score of 91.49% (the highest). Additionally, the inference speed is 4.54 times faster than the EM fusion algorithm with comparable accuracy, effectively balancing high precision and inference speed.

Key Words: annotation fusion; object detection; cow image dataset; weak supervision

0 引言

随着目标检测在自动驾驶、医学影像、智慧养殖等领域的应用广泛,高质量的标注数据便显得愈发重要^[1-6]。多个标注者之间的标注结果差异,例如位置偏差、漏标、误标等,会严重影响数据一致性与模型训练效果^[7]。多标注融合技术可通过集成多个模型或人工标注的结果,提升整

体标注质量^[8,9]。该技术可进一步细分为对象检测结果融合与多标注者融合,前者多用于融合多个模型输出的结果,利用各模型的预测互补性来提升检测性能;后者通过融合多个标注者的标注,得到更准确、一致的标注数据,以提升模型训练效率。

目前,常用的融合方法在效率、鲁棒性与复杂性之间难以兼顾。多数投票(Majority Voting, MV)方法简单直接,将大多数标注者认可的标签作为结果,多用于需要快

收稿日期:2026-01-15

作者简介:刘洋(1996-),男,华中农业大学信息学院硕士研究生,研究方向为计算机视觉;李国亮(1972-),男,博士,华中农业大学信息学院教授、博士生导师,研究方向为三维基因组学、农业人工智能。本文通讯作者:李国亮。

速融合的场景,但忽略了标注质量差异与空间定位精度。加权边界框融合(Weighted Boxes Fusion, WBF)方法利用所有检测框的置信度进行加权平均,以改善定位精度,但只关注高置信度的标注框,忽略了标注一致性的差异,即不同标注者或模型水平有所差异。高置信度但与群体共识偏离的标注框可能对融合结果产生影响,从而降低整体鲁棒性。期望最大化(Expectation-Maximization, EM)方法虽在弱监督或半监督目标检测场景中表现良好,能处理标签噪声,但计算复杂度较高、成本高昂,难以直接应用于大规模标注融合场景。贝叶斯组合检测(Bayesian Detector Combination, BDC)方法采用联合训练策略,可连接至任意目标检测模型,无需置信度信息,但需要与模型一起训练,无法直接在离线场景下进行多标注结果融合。

综上,如何在引入复杂模型训练或高昂计算成本的前提下,显式建模标注者一致性并提升融合结果的鲁棒性,是多标注目标检测融合中亟待解决的问题。为此,本文提出一种能在弱监督环境中提升标签质量的算法 CWF (Consistency-Weighted Fusion for Object Detection)。首先,引入一致性加权机制聚类同一目标的标注,并计算一致性得分;其次,在融合过程中根据一致性得分加权标注信息,以在保持融合准确性的同时提升时间效率。本文主要贡献为:①构建了牛只图像数据集领域内可用于标注融合评估的数据集,该数据集共有 4 000 张图像,每张图像既包含基准标注,又包含多人标注结果,平均每张图像由 8 人标注;②将一致性加权机制引入目标检测标注融合工作,以降低计算耗时,提升融合精度;③通过比较不同的标注融合算法,明晰了融合规模与标注性能之间的关系,对众包标注等实际应用具有指导意义。

1 相关工作

图 1 为标注融合工作的工作流程。首先,根据实际工作构建数据集,让不同标注者进行目标检测标注,或使用不同的目标检测模型进行检测;其次,通过不同的目标检测标注融合算法融合标注结果,得到准确性更高的标注结果,供下游任务使用。

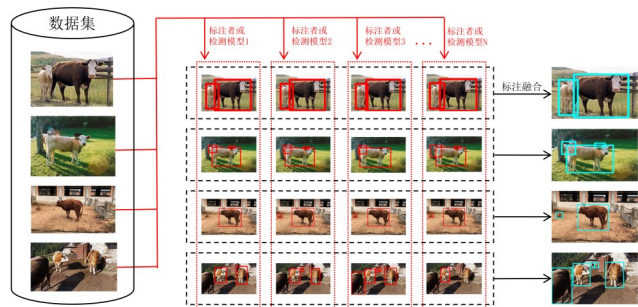


Fig. 1 Target detection annotation fusion process

图 1 目标检测标注融合流程

1.1 WBF

传统目标检测模型通常采用非极大值抑制(Non-Maximum Suppression, NMS)来筛选候选区域,但其存在因硬阈值而导致的目标丢失等问题^[10]。为此,研究人员提出一种改进的非极大值抑制算法(Soft-NMS),逐步降低与高分检测框重叠的低得分框的得分,而非直接移除这些框,从而减少因硬阈值导致的目标丢失^[11]。然而,上述算法的本质是从多个候选框中挑选出一个“认为的”最优候选框,并未真正融合多个候选框。

WBF 是目标检测标注融合领域最常见的方法,通过融合多个检测框的置信度来生成更准确的平均框,以保留更多细节信息,相较于 NMS、Soft-NMS 鲁棒性更强,在 Open Images 与 COCO 数据集上优势明显^[12]。Le 等^[13]也在医疗影像数据上证明了 WBF 算法的有效性,但 WBF 忽略了标注者之间的一致性,可能在融合少量错误标注时生成错误的融合结果。

1.2 基于聚类与 EM 的融合策略

WBF 的聚类式扩展包括 ALFA (Agglomerative Late Fusion Algorithm)。该方法通过层次聚类,对框进行聚合并加权,已在 PASCAL VOC2007、PASCAL VOC 2012 数据集上取得了优越的成绩^[14]。David 等^[15]提出的 EM 方法,被证明其融合效果优于多数投票算法^[16]。LAEM (Localization-Aware Expectation-Maximization) 借助重复标注的统计一致性,并融合期望最大化算法,在 VinDr-CXR Dataset (医疗影像数据集)、TexBiG dataset (文本数据集) 等数据集上的表现十分稳健^[17]。

1.3 贝叶斯模型

BDC (Bayesian Detector Combination) 使用贝叶斯框架建模标注者的类别可靠性与位置偏差,以检测器预测的类别概率与位置精度,联合推断真实框,实现标注融合与模型训练同步迭代。该方法不依赖准确的先验参数,相较于 WBF、EM 更自动化,在众包高噪声场景中适应性更高,可提升模型鲁棒性,在医疗影像数据集 VinDr-CXR Dataset 取得了较好的效果^[18]。

1.4 标注融合算法

现有研究多集中于多标注者、多类别的复杂场景,且大多数算法依赖标注者、模型置信度信息或需要对置信度进行建模,例如 WBF、EM、BDC 等。同时,现有研究侧重于提升检测精度,忽略了算法效率和复杂性。WBF 可融合多模型或多标注者的标注结果,但过于依赖置信度信息,忽略了标注者或模型的一致性差异,融合效果有待提升。基于 EM 的算法尽管在准确性上优势明显,但迭代计算复杂、耗时高。

为此,本文提出基于一致性加权的目标检测标注融合算法 CWF。该算法仅依赖不同标注者或模型的标注结果,无需额外的置信度信息,通过直接融合不同标注者或模型的标注结果,实现了在目标检测任务中的性能提升。

2 本文方法

图2为CWF算法整体框架。由此可见,算法旨在解决现实融合场景的3个核心挑战:标注者水平不均、缺乏真实值、高效离线计算,核心思想是评估每个标注者与群体共识的一致性。其中,一致性是指不同标注者或不同模型在同一目标上的标注结果彼此接近、相互支持的程度,强调的不是单个标注结果的置信度,而是不同标注者在同一目标上的空间重合度与群体共识强度。

本文认为一个标注者的权重应取决于其与群体共识的吻合程度。为此,设计了一个两阶段框架:首先,基于IoU对所有候选框进行分组,每个组内所有候选框构成一个簇;其次,在此基础上构建群体共识框,以共识框为暂时标准,计算各标注者与之匹配的精度与召回率,并将其调和平均数作为一致性权重量化指标。该设计使CWF无需依赖先验的置信度或二次训练模型,即可显式建模并利用标注者差异,在抑制异常标注影响的同时,保留可能有效但权重较低的目标。

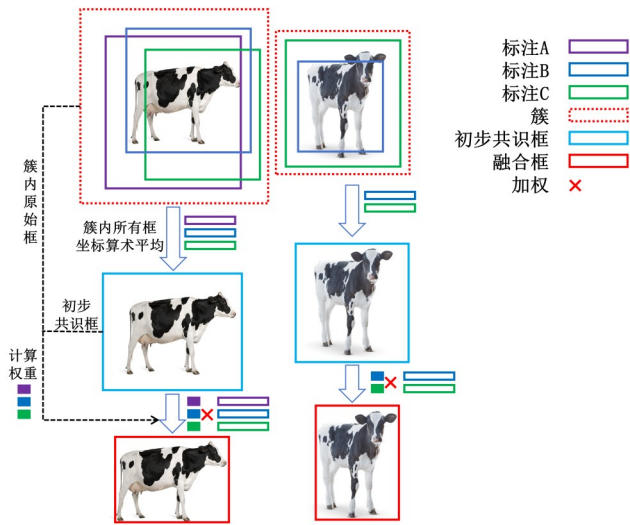


Fig. 2 Overall framework of the proposed CWF algorithm

图2 CWF算法整体框架

2.1 问题描述与符号约定

假设每张图有 M 位标注者 (Annotators) 独立标注,第 j 位标注者给出的标注集合 $\mathcal{A}^{(j)}$ 为:

$$\mathcal{A}^{(j)} = \{b_1^{(j)}, b_2^{(j)}, \dots, b_{n_j}^{(j)}\} \quad (1)$$

式中:单个框以矩形坐标表示 $b = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ 。

每个标注者标注的框数量不固定,用一个自然数 n_j 表示。所有标注者对单张图像的标注集合 $\mathcal{B}$ 为:

$$\mathcal{B} = \bigcup_{j=1}^M \mathcal{A}^{(j)} \quad (2)$$

目的是从标注集合 $\mathcal{B}$ 中生成融合后的标注集合 $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$,使 $\mathcal{F}$ 与真实标注 (Ground Truth) 在位置和大小上尽可能接近,同时在计算时间上具有优势。

为了后文表述方便,定义两个框 P, Q 的交并比 (intersection over union, IoU) 为:

$$\text{IoU}(P, Q) = \frac{\text{area}(P \cap Q)}{\text{area}(P \cup Q)} \quad (3)$$

2.2 CWF算法算法步骤

算法1 CWF算法

输入:多个目标检测标注文件合集 $\mathcal{A}^{(j)}$ 。

输出:每张图像的融合框集合 $\mathcal{F}$ 。

1. 解析标注文件,用式(1-2)提取单张图像所有标注者的边界框的集合 $\mathcal{B}$
2. 用式(3)计算集合 $\mathcal{B}$ 内候选框的IoU,通过式(5)将候选框逐步分组为若干簇 C_k
3. FOR 每个簇 DO
4. 用式(7-8)计算初步共识框 c_k
5. END
6. 基于所有初步共识框合集,用式(9-12)计算每个标注者的权重 $\tilde{\omega}_j$
7. FOR 每个簇 DO
8. 用式(13-16)计算融合框 f_k
9. END
10. 输出融合框 f_k 的集合 $\mathcal{F}$

2.3 初步共识框与标注者一致性度量

基于IoU阈值 τ_{match} 定义IoU的邻接关系。

$$b_p \sim b_q \Leftrightarrow \text{IoU}(b_p, b_q) \geq \tau_{\text{match}} \quad (4)$$

通过邻接关系将集合 $\mathcal{B}$ 中的所有候选框逐步分组为 K 个簇 C_k 。每个簇内的候选框数量不定,但应是一个正整数,如式(6)所示。

$$C_k = \{b \in \mathcal{B} | \exists b_1, b_2, \dots, b_r \in \mathcal{B}, b \sim b_1 \sim b_2 \sim \dots \sim b_r\} \quad (5)$$

$$\mathcal{B} = \bigcup_{k=1}^K C_k, C_k \cap C_l = \emptyset, k \neq l \quad (6)$$

每个簇内的候选框代表同一个潜在的检测目标。对于簇 C_k 中的若干框,先构建初步共识框 c_k ,默认采用坐标算术平均进行计算。

$$c_k = (\bar{x}_{\min}, \bar{y}_{\min}, \bar{x}_{\max}, \bar{y}_{\max}) \quad (7)$$

式中: $\bar{x}_{\min}$ 的计算公式由式(8)所示, $\bar{y}_{\min}, \bar{x}_{\max}, \bar{y}_{\max}$ 同理。

$$\bar{x}_{\min} = \frac{1}{|C_k|} \sum_{b \in C_k} x_{\min}(b) \quad (8)$$

本文以一张图里所有初步共识框集合 $\{c_k\}_{k=1}^K$ 为参考,对每位标注者 j 计算一致性分数。 P_j, R_j 、权重 ω_j 的计算公式如式(9)–式(11)所示。

$$P_j = \begin{cases} \frac{|\{b: \exists b \in \mathcal{A}^{(j)}, \text{IoU}(b, c_k) \geq \tau_{\text{match}}\}|}{|\mathcal{A}^{(j)}|}, & |\mathcal{A}^{(j)}| > 0, \\ 0, & |\mathcal{A}^{(j)}| = 0. \end{cases} \quad (9)$$

$$R_j = \frac{|\{c_k: \exists b \in \mathcal{A}^{(j)}, \text{IoU}(b, c_k) \geq \tau_{\text{match}}\}|}{K} \quad (10)$$

$$\omega_j = \begin{cases} \frac{2P_j R_j}{P_j + R_j}, & P_j + R_j > 0, \\ 0, & \text{其它} \end{cases} \quad (11)$$

式中: P_j 表示该标注者的标注结果中有多少能匹配到初步共识框; R_j 表示初步共识框有多少被该标注者的标注结果匹配; ω_j 为 P_j 、 R_j 的调和平均; τ_{match} 表示判断两个检测框是否匹配的 IoU 阈值, 本文在初步共识框计算阶段、标注结果与共识框的匹配阶段, 均设置统一的 $\tau_{\text{match}} = 0.5$ (参考目标检测领域的通用做法)。

上述过程通过集体共识评估每个标注者的可靠性与一致性, 并以此为每个标注者赋予不同权重。

随后, 对所有 ω_j 进行归一化:

$$\tilde{\omega}_j = \frac{\omega_j}{\sum_{i=1}^M \omega_i} \quad (12)$$

2.4 基于一致性加权的检测标注融合

对于簇 C_k , 簇内的每个原始框 b 来自标注者 $j(b)$ 。 C_k 的融合框 f_k 的坐标按照该标注者权重的加权平均得到, 如式(13)~式(16)所示。

$$x_{\min}(f_k) = \frac{\sum_{b \in C_k} \tilde{\omega}_{j(b)} \cdot x_{\min}(b)}{\sum_{b \in C_k} \tilde{\omega}_{j(b)}} \quad (13)$$

$$y_{\min}(f_k) = \frac{\sum_{b \in C_k} \tilde{\omega}_{j(b)} \cdot y_{\min}(b)}{\sum_{b \in C_k} \tilde{\omega}_{j(b)}} \quad (14)$$

$$x_{\max}(f_k) = \frac{\sum_{b \in C_k} \tilde{\omega}_{j(b)} \cdot x_{\max}(b)}{\sum_{b \in C_k} \tilde{\omega}_{j(b)}} \quad (15)$$

$$y_{\max}(f_k) = \frac{\sum_{b \in C_k} \tilde{\omega}_{j(b)} \cdot y_{\max}(b)}{\sum_{b \in C_k} \tilde{\omega}_{j(b)}} \quad (16)$$

3 实验结果与分析

3.1 实验环境

本文实验包含两个阶段: 目标检测模型推理阶段和标注融合算法评估阶段。两阶段使用的软、硬件环境并不相同, 具体配置如表1所示。

Table 1 Experimental configuration

表1 实验配置

参数	型号版本
实验阶段	标注融合阶段
系统	Windows 11
CPU	Intel Core i5-13500H
GPU	未使用
Python	3.6.13
实验阶段	目标检测模型推理阶段
系统	CentOS Linux 7
CPU	Intel Xeon Gold 6240
GPU	NVIDIA A100(40 GB)
CUDA Version	12.1
Python	3.9.20
PyTorch	2.5.1

3.2 数据集与标注

本文从14个数据集中人工选取含有牛只的图像, 构建了包含4 000图像的自建牛只图像数据集, 如表2所示。

Table 2 Source and quantity of cow images

表2 牛只图像来源与数量

数据集	牛只图像数量
COCO Dataset	248
PASCAL VOC	23
Google Open Images Dataset	378
Animals-10	247
Animal-Pose Dataset	27
Ap-10k	27
Animal Image Dataset (90 Different Animals)	6
Holstein Cattle Recognition	168
Cattle Body Parts Dataset for Object Detection Dataset	63
Cow-images	32
Cows2021 Database	1368
CBVD-5	647
Dairy-cow	510
NWAFU Cattle Dataset	256

由表2可见, 14个数据集中包括通用大型数据集7个, 包括COCO Dataset、PASCAL VOC、Google Open Images Dataset、Animals-10、Animal-Pose Dataset、Ap-10k、Animal Image Dataset (90 Different Animals)^[19-25]; 专门的牛只图像数据集7个, 包括Holstein Cattle Recognition、Cattle Body Parts Dataset for Object Detection Dataset、Cow-images、Cows2021 Database、CBVD-5、Dairy-cow、NWAFU Cattle Dataset^[26-32]。

本文将每100张图像打包为一份, 共40个包, 每个包里的100张图像由4~11人独立标注, 平均每个包由8人标注, 具体情况如图3所示。其中, 标注内容为单一类别“cow”, 标注文件以VOC XML格式保存, 标注工具使用LabelImg。数据集包含以下情形: 单个牛只、多个牛只(有重叠与不重叠)、无牛只, 牛只包含不同的尺度、光影、角度, 如图4所示。实验基准标注数据由经验丰富的标注专家完成, 并以其作为基准来评估不同算法融合标注的性能。实验中, 4 000张图像均被用作融合标注和性能评估。

3.3 评价指标

实验采用精确率(Precision, P)、召回率(Recall, R)、F1分数(F1-score, F1)等指标衡量模型性能, 各个评价指标的取值区间为[0, 1], 计算公式如式(17)~(19)所示。

$$P = \frac{TP}{TP + FP} \quad (17)$$

$$R = \frac{TP}{TP + FN} \quad (18)$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \quad (19)$$

式中: TP表示实际为牛只且被模型预测为牛只的实例; FP表示实际为非牛只但模型预测为牛只的实例; FN表示实际为牛只但是模型预测为非牛只的实例。

同时, 计算宏平均(Macro)和微平均(Micro)两种指标。

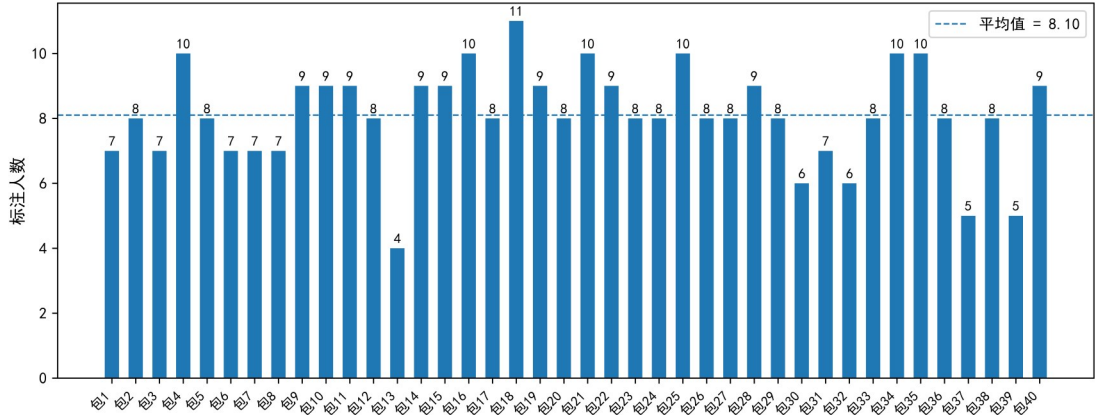


Fig. 3 Distribution of annotated people

图3 标注人数分布情况



Fig. 4 Partial cow images

图4 部分牛只图像

宏平均指标是先计算每张图像的精度与召回率,再对所有图像取平均,能保证即使图像中的目标数不同,也能使每张图像的权重相等;微平均指标是将所有图像的 TP、FP、FN 累加起来后再计算精度与召回率,使每个目标的检测结果都直接参与了最终统计。

时间指标用 ms/img 和 ms/box 进行衡量。其中,ms/img 表示每张图像的平均融合时间。考虑到不同的图像是由不同数量的标注者完成,每个标注者标注的目标数量也存

在差异,为了尽可能消除人数差异带来的影响,本文提出 ms/box 指标来表示平均每个检测框的耗时,box 为所有标注者提供的原始框总数。

3.4 比较方法

基准(Baseline):不做任何处理,直接计算每个标注员的标注结果与基准标注的 F1,整体性能通过 F1 平均值进行衡量。交集法融合:初始以第一个标注员的结果作为融合结果,后续依次与新标注员结果进行比较:若当前融合框与新标注员框 $IoU \geq 0.5$,则取二者交集作为更新后的融合框;否则保留原框,同时将新框加入。WBF 法融合:与文献[12]的流程一致,由于没有每位标注者的置信度,实际计算过程中置信度均设为 1。EM 法融合:基于期望最大化的融合算法,分为 E 步和 M 步,直到达到最大迭代次数(10)或者达到收敛阈值(0.001)时停止。

3.5 实验结果

3.5.1 最佳融合规模

最优融合规模对实际目标检测应用或众包标注人员配备具有指导意义,不同融合算法下 Macro/Micro-F1 的最优次数分布情况如图 5 所示。

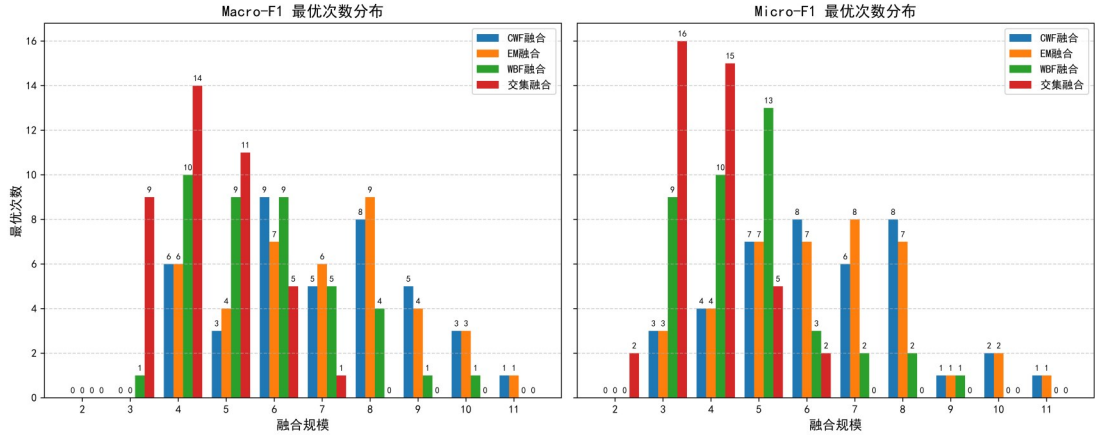


Fig. 5 Optimal frequency distribution of Macro/Micro-F1 under different fusion algorithms

图5 不同融合算法下 Macro/Micro-F1 的最优次数分布情况

由图5可见,针对交集、WBF等较简单的算法,Macro-F1最优次数集中在3~6人, Micro-F1最优次数集中在3~5人;针对CWF、EM等较复杂的算法,Macro-F1最优次数集中在6~8人, Micro-F1最优次数集中在5~8人。

综上,在实际应用中,应该根据融合算法的复杂程度,控制模型融合规模或每张图像的标注人数保持在5人左右,与文献[18]的结论基本一致。

3.5.2 τ_{match} 超参数敏感性分析

为分析超参数 τ_{match} 对标注融合性能的影响,在保持其他参数不变的条件下,将 τ_{match} 分别设置为0.3、0.4、0.5、0.6和0.7,并在相同数据集上进行实验,如表3、图6所示。

Table 3 Performance of CWF algorithms under different τ_{match}

表3 不同 τ_{match} 下CWF算法的性能 (%)

τ_{match}	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1
0.3	96.90	87.73	92.05	93.53	74.95	83.07
0.4	95.59	89.28	92.29	90.77	77.91	83.67
0.5	93.37	89.80	91.49	85.76	78.75	81.86
0.6	89.46	89.41	89.34	77.64	77.80	77.34
0.7	82.58	88.06	85.07	65.21	74.87	69.12

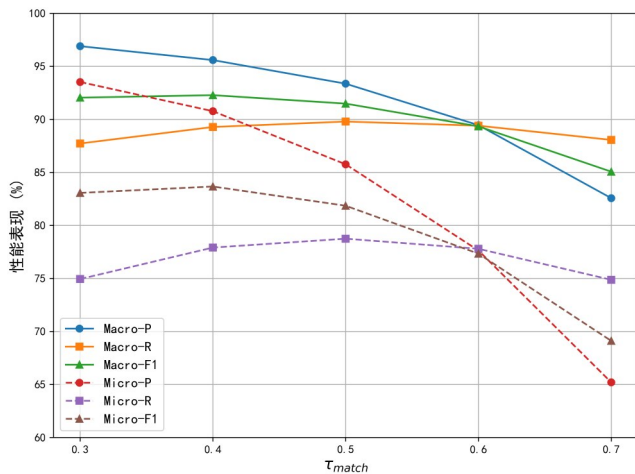


Fig. 6 τ_{match} -performance index variation curve

图6 τ_{match} -性能指标变化曲线

实验表明,随着 τ_{match} 增大,CWF算法的Precision呈下降趋势,Recall相对稳定。当 τ_{match} 位于0.3~0.5区间时,各项性能指标变化较平稳,融合结果对 τ_{match} 的取值不敏感,说明算法在该匹配阈值区间下具有较好的鲁棒性。当 $\tau_{match} > 0.6$ 时,性能下降明显。在综合考虑Precision、Recall的变化情况后,决定在后续实验中采用 $\tau_{match} = 0.5$ 作为默认设置。

3.5.3 不同融合算法的性能

不同目标检测标注融合算法在性能上各有优劣,本文将交集法融合、WBF法融合、EM法融合和CWF法融合进行比较实验,如表4所示。由此可知,相较于基准,4种标注融合算法的性能均有提升,CWF算法在Macro-F1、Micro-F1上分别提升8.55%和9.42%。总体而言,CWF算法在精度与召回表现上最为突出,Macro-F1、Micro-F1最高,

证明了算法在目标检测标注融合任务中的有效性。

Table 4 Performance of different fusion algorithms

表4 不同融合算法的性能

融合算法	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1
基准	92.84	75.38	82.94	93.51	60.31	72.44
交集	90.39	89.97	90.06	79.82	79.27	79.04
WBF	91.87	90.21	90.94	82.68	79.74	80.79
EM	93.36	89.79	91.48	85.75	78.74	81.85
CWF	93.37	89.80	91.49	85.76	78.75	81.86

随着融合规模增加,不同融合算法的性能也会发生变化,如图7所示。在Macro-F1方面,交集、WBF、EM、CWF等4种算法分别在融合规模为4、6、6、7时最优,随后开始下降;在Micro-F1方面,交集、WBF、EM、CWF等4种算法分别在融合规模为3、5、7、7时达到最优,随后开始下降,且交集法下降最明显。

在融合规模为10、11时,虽然F1值有所上升,但由于参与计算的包数太少,存在偶然性。由图8可知,在不同标注融合算法下,超过一定阈值后,40个包的个体趋势仍然趋于平稳或下降,仅在融合规模为10、11时,均值因样本数量太少而上升。综上,当融合规模达到一定阈值后,继续增加融合规模反而会导致融合结果变差,这意味着在实际应用中,并非标注人员或模型规模越大越好。

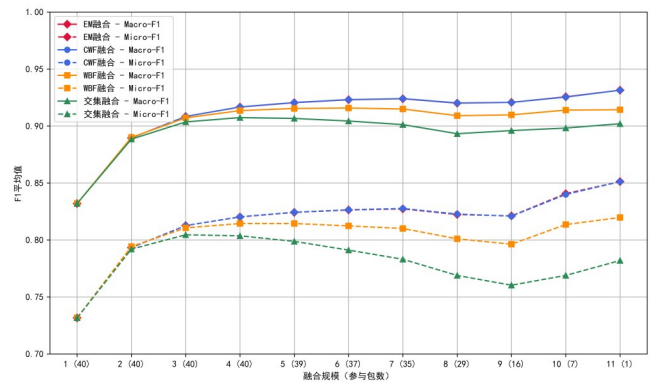


Fig. 7 Macro/Micro-F1 of different fusion algorithms at different fusion scales (multiple annotators)

图7 不同融合算法在各融合规模下的Macro/Micro-F1(多个标注者)

3.5.4 不同融合算法的运行效率

在计算耗时上,不同标注融合算法差异更加明显。表5为不同标注融合算法的耗时情况,分别统计了平均每张图像的处理时间(ms/img)以及平均每个标注框的处理时间(ms/box)。由此可知,交集法与WBF法在效率上最为突出,平均每张图像处理时间仅为71.27 ms与80.32ms,单框耗时分别为2.61 ms与2.93 ms,适合对实时性要求较高的场景。EM算法虽然在精度指标上表现较好,但运行效率显著低于其他方法,平均耗时达到1 010.84 ms/img(36.75 ms/box),限制了其在大规模应用中的可行性。CWF算法运行效率远优于EM算法,平均每张图像的处理时间为222.74 ms,相较于EM算法快4.54倍,平均每个检测框的处

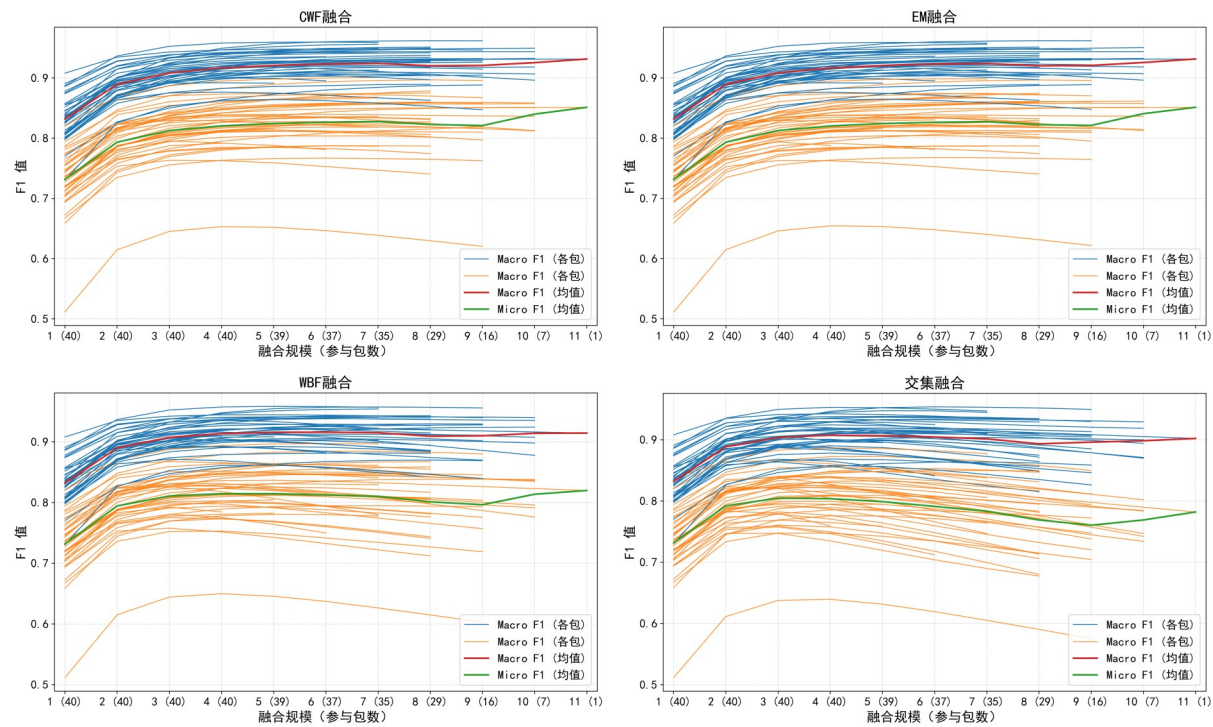


Fig. 8 Changes in packages and global Macro/Micro-F1 in different models

图 8 不同模型中各包与全局 Macro/Micro-F1 的变化情况

理时间为 7.99 ms,相较于 EM 算法快 4.60 倍。综上,相较于 WBF 算法和 EM 算法,CWF 算法在精度与效率之间实现了较好的平衡。

Table 5 Computational cost of different fusion algorithms

表 5 不同融合算法的耗时		
融合算法	ms/img	ms/box
交集	71.24	2.61
WBF	80.32	2.93
EM	1 010.84	36.75
CWF	222.74	7.99

3.5.5 结果可视化

图 9 为 4 种不同融合算法的融合结果,上半部分用于比较 4 种融合算法、不同标注员(红色框)及真值(蓝色框),下半部分单独展示了 CWF、EM、WBF、交集算法融合后的检测框与真值的区别。由此可见,交集算法的性能最差,出现了冗余框重复框选同一目标的情况;WBF 算法的

抗噪声能力较弱,检测框偏大;EM、CWF 算法的融合效果优秀,性能接近。

3.5.6 融合不同目标检测模型的结果

考虑到上述实验中的 40 个包的标注人数并不一致,为了检验本文算法对目标检测模型输出结果的可用性,用 6 种目标检测模型检测 4 000 张牛只图像,并使用不同融合算法对 6 种目标检测模型的结果进行融合,结果如表 6 所示。由此可知,CWF 算法与 EM 算法性能接近;相较于 6 种模型的 F1 均值,CWF 算法的 Macro-F1 提升 3.29%,Micro-F1 提升 3.1%。

图 10 为 6 种目标检测模型的检测结果在不同融合算法下,随着融合检测结果数量增加时的 F1 分数。由此可见,交集、WBF、EM、CWF 等 4 种算法的 Macro-F1 分别在融合规模为 3、3、4、4 时达到最优,随后开始下降;Micro-F1 分别在融合规模为 2、2、3、3 时达到最优,随后开始下降。

Table 6 Performance of different fusion algorithms in fusing the results of six object detection models

表 6 不同融合算法对 6 种目标检测模型的结果进行融合的性能 (%)

模型/算法	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1
Yolov5su	74.84	56.95	60.88	82.77	39.52	53.50
Yolov8s	71.32	54.56	58.13	80.99	39.14	52.78
Yolov9s	72.27	56.13	59.38	83.04	40.81	54.73
Yolov10s	74.06	61.11	63.63	84.07	45.19	58.79
Retinanet	69.19	64.76	62.30	57.50	47.02	51.73
Maskrcnn	63.43	70.93	62.08	52.68	53.54	53.10
均值	70.85	60.74	61.07	73.51	44.20	54.11
交集	55.80	75.31	59.31	42.60	62.87	50.79
WBF	60.34	76.26	62.73	48.37	63.84	55.04
EM	64.26	73.01	64.21	53.86	58.85	56.24
CWF	64.44	73.17	64.36	53.82	58.80	56.20

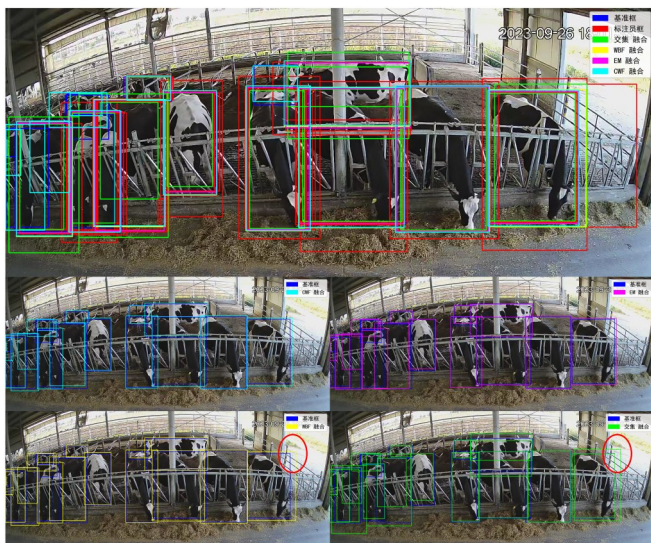


Fig. 9 Visual comparison of fusion results obtained by different methods (red ellipses indicate regions with noticeable differences)

图9 不同融合算法结果的可视化对比(红色椭圆表示差异区域)

再次证明了,在目标检测标注融合的过程中,融合规模与融合效果并非呈正相关。

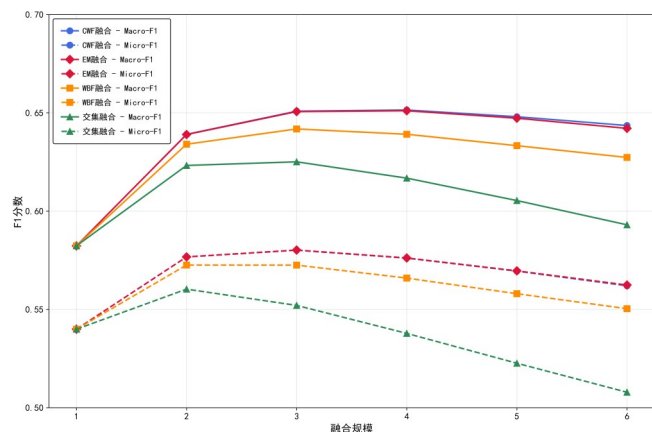


Fig. 10 Macro/Micro-F1 of different fusion algorithms at different fusion scales (multiple models)

图10 不同融合算法在各融合规模下的Macro/Micro-F1(多个模型)

4 结语

在目标检测任务中,高质量的标注数据对训练任务非常重要。为了得到高质量的标注数据,通常采用多人标注,再将标注结果融合的策略,但未充分考虑算法运行效率与复杂性之间的平衡。为此,本文提出CWF算法,在缺乏真实值时初步计算群体共识框,并基于共识框计算每个标注者的水平,为每个标注者赋予权重,最终基于原始框与对应的权重计算融合框。

本文方法将“标注可靠性”的来源从单个标注的置信度,转为标注者与群体共识的一致性,并显式建模这种一致性,从而在抑制离群噪声的同时,保留了真实但不完全一致的标注。实验表明,所提方面相较于基准、交集算法、

WB、EM算法,在F1值与最优算法EM算法持平的同时,减少了计算成本。

比较实验可知,传统交集算法(71.24 ms/img)、WB算法(80.32 ms/img)在计算成本上更具优势,但Macro-F1较低(分别为90.06%和90.44%);EM算法虽可达到更高的Macro-F1(91.48%),但计算成本却高达1 010.84 ms/img,难以满足实际应用需要;CWF算法在保持高Macro-F1(91.49%)的同时,计算耗时减少到了222.74 ms/img,兼顾了准确性与运行效率的需要,凸显了其在算力受限场景下的应用价值。

更重要的是,CWF算法无需要标注结果的置信度信息,便可在众包标注等信息受限的场景下,融合多位标注者的标注结果,以获得更好的融合标注结果。此外,该算法也可与多个目标检测模型的检测结果进行融合,以进一步提升模型检测精度。具体地,相较于6个目标检测模型的检测结果均值,CWF算法的Macro-F1值提升3.29%,Micro-F1值提升3.1%。

本文通过比较不同算法在不同融合规模下的性能发现,不同融合算法虽在具体分布上略有差异,但Macro-F1的最优融合规模集中在4~6, Micro-F1集中在3~5,且进一步增加融合规模会导致性能下降。因此,在实际应用中,应将模型融合规模或每张图像的标注人数控制在5左右,以保证融合效果最优。

然而,当前研究仍存在一定的局限性:实验所用数据集仅包含单一类别的标注,缺乏多类别场景下的验证。未来,尝试将算法改良为适用于多类别目标检测的标注融合,以提升算法在复杂场景下的性能。

参考文献:

- [1] CARRANZA-GARCÍA M, LARA-BENÍTEZ P, GARCÍA-GUTIÉRREZ J, et al. Enhancing object detection for autonomous driving by optimizing anchor generation and addressing class imbalance [J]. Neurocomputing, 2021, 449: 229-244.
- [2] CONTRERAS M, JAIN A, BHATT N P, et al. A survey on 3D object detection in real time for autonomous driving [J]. Frontiers in Robotics and AI, 2024, 11: 1212070.
- [3] ALBUQUERQUE C, HENRIQUES R, CASTELLI M. Deep learning-based object detection algorithms in medical imaging: Systematic review [J]. Heliyon, 2025, 11(1): 41137.
- [4] ELHANASHI A, SAPONARA S, ZHENG Q, et al. AI-Powered object detection in radiology: current models, challenges, and future direction [J]. Journal of Imaging, 2025, 11(5): 141.
- [5] SHAO D, HE Z, FAN H, et al. Detection of cattle key parts based on the improved Yolov5 algorithm [J]. Agriculture, 2023, 13(6): 1110.
- [6] HAO W, REN C, HAN M, et al. Cattle body detection based on YOLOv5-EMA for precision livestock farming [J]. Animals, 2023, 13(22): 3535.
- [7] MURRUGARRA-LLERENA J, KIRSTEN L, JUNG C R. Can we trust bounding box annotations for object detection? [C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2022:

- 4812–4821.
- [8] HAN R, WANG C, WANG Y, et al. Defect detection in EBSM components through selective box fusion of modern object detection[J]. Scientific Reports, 2025, 15(1): 11899.
- [9] RAMAKRISHNA A, GUPTA R, GROSSMAN R B, et al. An expectation maximization approach to joint modeling of multidimensional ratings derived from multiple annotators[C]// Interspeech, 2016: 1555–1559.
- [10] GOOL L V, NEUBECK A. Efficient non-maximum suppression [C]// 18th International Conference on Pattern Recognition, 2006: 850–855.
- [11] BODLA N, SINGH B, CHELLAPPA R, et al. Soft-NMS — improving object detection with one line of code[C]// 2017 IEEE/CVF International Conference on Computer Vision, 2017: 5562–5570.
- [12] ROMAN S, WEIMIN W, TATIANA G. Weighted boxes fusion: ensembling boxes from different object detection models[J]. Image and Vision Computing, 2021, 107: 104117.
- [13] LE K H, TRAN T V, PHAM H H, et al. Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis [J]. IEEE Access, 2023, 11: 14105–14114.
- [14] RAZINKOV E, SAVELEVA I, MATAS J. ALFA: agglomerative late fusion algorithm for object detection [C]// 2018 IEEE/CVF International Conference on Pattern Recognition, 2018: 2594–2599.
- [15] DAWID A P, SKENE A M. Maximum likelihood estimation of observer error-rates using the em algorithm[J]. Journal of the Royal Statistical Society: Series C(Applied Statistics), 1979, 28(1): 20–28.
- [16] RAYKAR V C, YU S, ZHAO L H, et al. Supervised learning from multiple experts: whom to trust when everyone lies a bit [C]// Proceedings of the 26th Annual International Conference on Machine Learning, 2009: 889–896.
- [17] TSCHIRSCHWITZ D, BENZ C, FLOREK M, et al. Drawing the same bounding box twice? coping noisy annotations in object detection with repeated labels[C]// DAGM German Conference on Pattern Recognition, 2023: 605–623.
- [18] TAN Z Q, ISUPOVA O, CARNEIRO G, et al. Bayesian detector combination for object detection with crowdsourced annotations [C]// European Conference on Computer Vision, 2025: 329–349.
- [19] LIN T-Y, MAIRE M, BELONGIE S J, et al. Microsoft COCO: common objects in context[DB/OL]. <https://arxiv.org/abs/1405.0312>.
- [20] EVERINGHAM M, ESLAMI S M A, GOOL L, et al. The pascal visual object classes challenge: a retrospective [J]. International Journal of Computer Vision, 2015, 111(1): 98–136.
- [21] KUZNETSOVA A, ROM H, ALLDRIN N, et al. The open images dataset V4 [J]. International Journal of Computer Vision, 2020, 128(7): 1956–1981.
- [22] ALESSIO C. Animals-10 [EB/OL]. <https://www.kaggle.com/datasets/alessiocorrado99/animals10/>.
- [23] CAO J, TANG H, FANG H, et al. Cross-domain adaptation for animal pose estimation [DB/OL]. <https://arxiv.org/abs/1908.05806>.
- [24] YU H, XU Y, ZHANG J, et al. Ap-10k: a benchmark for animal pose estimation in the wild [DB/OL]. <https://arxiv.org/abs/2108.12617>.
- [25] BANERJEE S. Animal image dataset (90 different animals) [EB/OL]. <https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals>.
- [26] BHOLE A, FALZON O, BIEHL M, et al. A computer vision pipeline that uses thermal and RGB images for the recognition of Holstein cattle [C]// International Conference on Computer Analysis of Images and Patterns, 2019: 108–119.
- [27] KHALILI A. Cattle body parts dataset for object detection [EB/OL]. <https://www.kaggle.com/datasets/alikhalilit98/cattle-body-parts-dataset-for-object-detection>.
- [28] AFNAN. Cow-images [EB/OL]. <https://www.kaggle.com/datasets/afnanamin/cow-images>.
- [29] GAO J, BURGHARDT T, ANDREW W, et al. Towards self-supervision for video identification of individual holstein-friesian cattle: The cows2021 dataset [DB/OL]. <https://arxiv.org/abs/2105.01938>.
- [30] LI K, FAN D, WU H, et al. A new dataset for video-based cow behavior recognition [J]. Scientific Reports, 2024, 14(1): 18702.
- [31] TWISDU. Dairy cow [EB/OL]. <https://www.kaggle.com/datasets/twisdu/dairy-cow>.
- [32] FAN Q, LIU S, LI S, et al. Bottom-up cattle pose estimation via concise multi-branch network [J]. Computers and Electronics in Agriculture, 2023, 211: 107945.

(责任编辑:刘嘉文)