中国科学院大学
University of Chinese Academy of Sciences

智能移动机器人结题报告

DETR调研实施报告

Intelligent Mobile Robot Project Proposal: DETR Research Implementation Report

Code：https://github.com/ZhijiangTang/Endterm-Report

汇报人：唐治江

2025年4月5日

中国科学院大学
University of Chinese Academy of Sciences

目录

CONTENTS

1

DETR概述

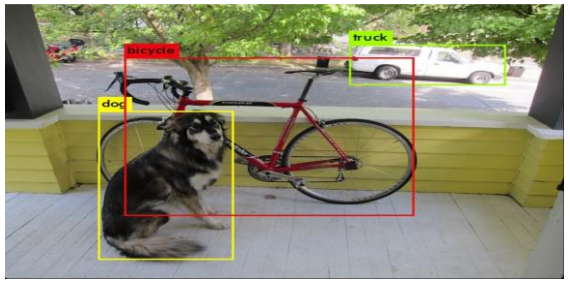
DETR Overview

项目代码

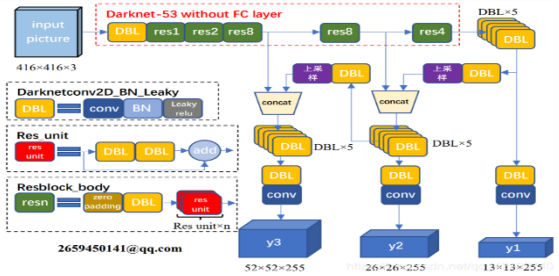
01

目标检测

在图像或视频中自动识别并定位出所有感兴趣的目标对象。两阶段目标检测方法通过先生成候选区域，再对每个区域进行精细分类和边界框回归。YOLO（You Only Look Once）是一种One-Stage目标检测算法，通过单次前向传播将图像划分成网格，直接预测目标的边界框和类别。



目标检测



YOLO V3架构图

02

DEtection TRansformer

Facebook提出DEtection TRansformer直接输出目标的边界框和类别，省去了传统目标检测NMS和锚框Anchor。

Deformable DETR引入可变形注意力机制，有效加速了模型的收敛。

RT-DETR第一个实时的端到端DETR。

2020 ECCV

2021 ICLR

2024 CVPR

2025 CVPR

Facebook提出**DE**tection **TR**ansformer直接输出目标的边界框和类别，省去了传统目标检测NMS和锚框Anchor。

DEIM在原始 DETR 框架上，通过改进预测与真实目标匹配机制来加速模型收敛的一种变体。

1.2 DETR与YOLO对比分析

项目代码

	YOLO	DETR End-to-End Object Detection with Transformers
基本架构	**CNN为主**，基于锚框和网格预测。	**Transformer**编码器-解码器，**无锚框机制。**
检测方式	端到端目标检测，基于回归，需要 **NMS** 进行后处理	端到端目标检测，**集合匹配**
训练收敛	**训练快**（100+ epoch），显存友好	**收敛慢**（需500+ epoch），显存消耗高
适合任务	**轻量化部署场景**，实时性要求高的场景	适用于**高精度检测任务**，如自动驾驶

2

DETR相关论文分析

Analysis of DETR-related papers

2.1 DEtection TRansformer

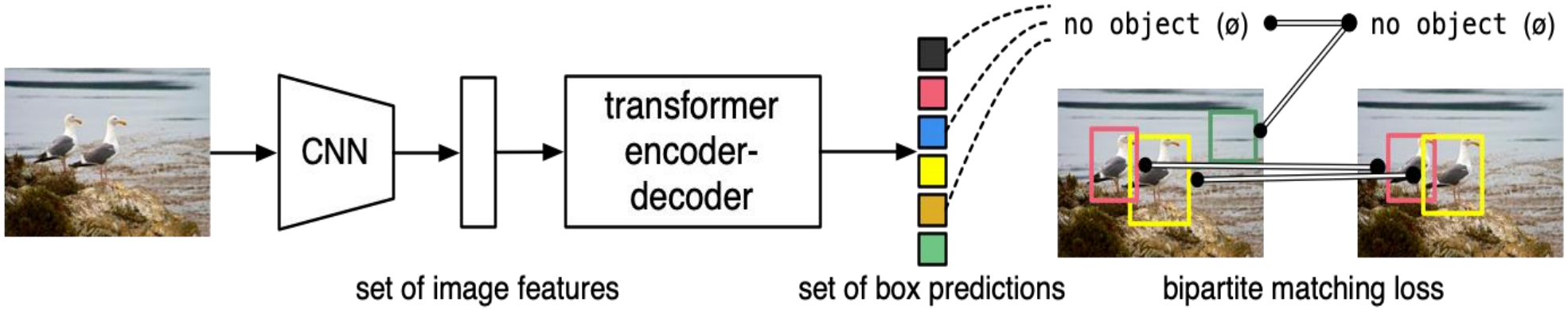
项目代码

DETR核心结构

DETR（DEtection TRansformer）的核心结构是结合**CNN提取特征的ResNet主干**、**Transformer编码全局关系**，并用**匈牙利匹配和Bipartite Loss**进行端到端目标检测。

Transformer的作用

Transformer 在 DETR 中的作用是建模全局关系，利用自注意力机制捕捉目标间的长距离依赖，并通过**查询向量（object queries）**生成最终检测结果。



DETR 通过将常见的 CNN 与 Transformer 架构相结合来直接（并行）预测最终的检测集。

2.1匈牙利匹配

算法

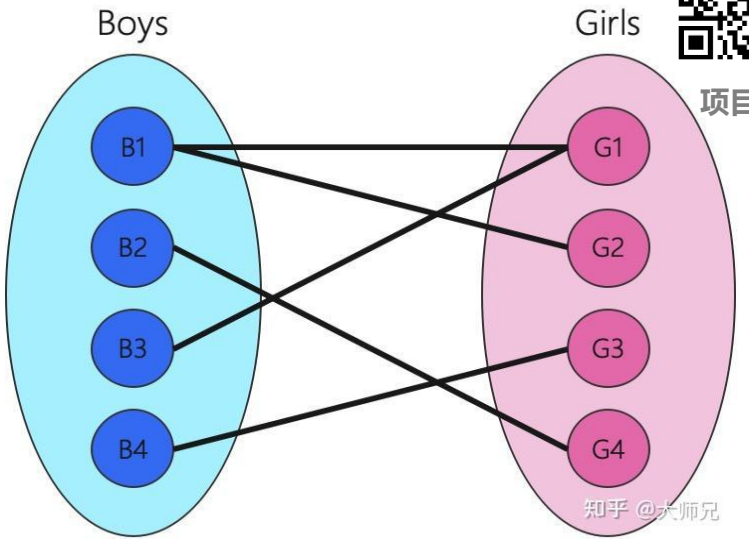
问题：匈牙利算法是一种用于求解二分图最优匹配的多项式时间算法。**E.g.** 包含男孩和女孩，连线代表暧昧关系，怎么凑成最多的情侣数?

算法：构造代价矩阵 → 行列归一化 → **寻找零覆盖匹配** → 若未匹配完则调整矩阵 → 迭代直至找到最优匹配。

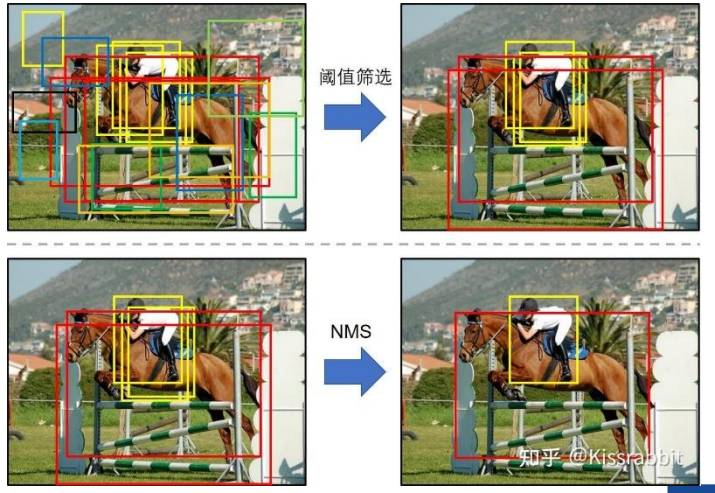
优势

- 端到端检测，**无需 NMS**
- 全局最优匹配，提升检测质量
- **无需手动设计 Anchor**，适应性更强
- Transformer **处理长距离依赖**，增强复杂场景理解

项目代码



Boys集合有4个男孩，Girls集合也有4个女孩，连线代表它们可以凑成一对情侣



YOLO的两套阈值

2.2 Real Time DEtection TRansformer

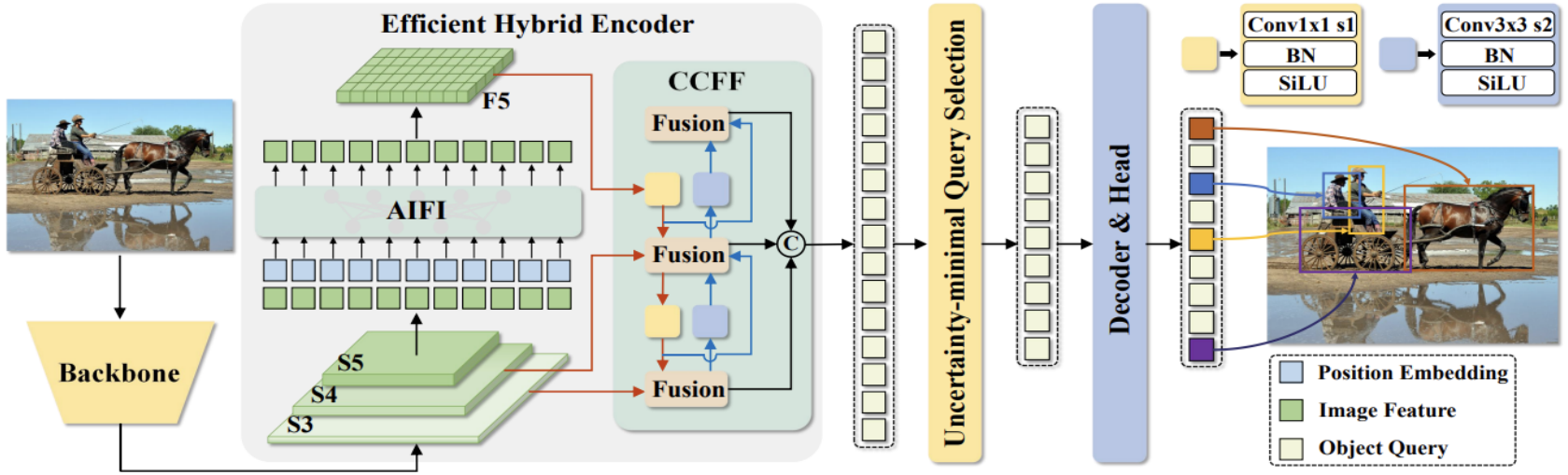
项目代码

RT-DETR核心结构

RT-DETR（Real-Time Detection Transformer）在 DETR 的基础上**优化解码结构、引入高效注意力机制**，并结合 CNN 特征提取，以提升推理速度，实现端到端实时目标检测。

时间优势

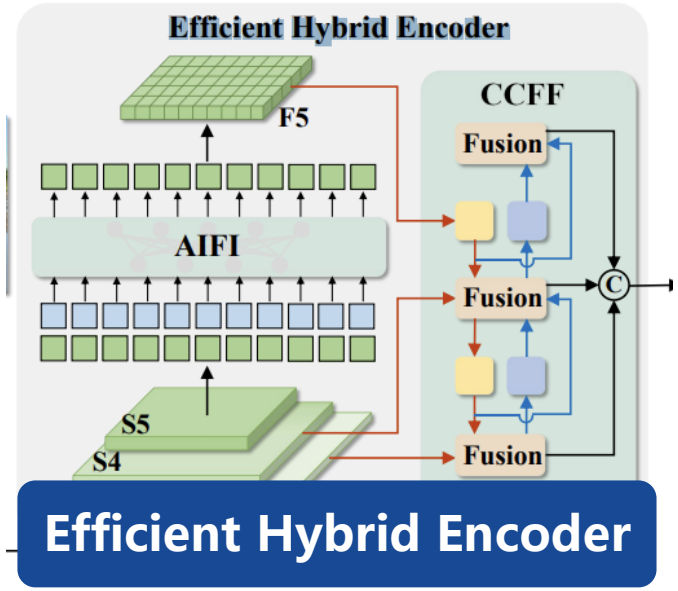
使用 CNN 进行**多层特征提取**，减少**一层Transformer** 直接处理高分辨率特征图的计算量，提升推理速度。**Uncertainty-minimal Query Selection**：优先选择即置信度最高的查询向量来进行最终目标框预测。



RT-DETR网络结构图

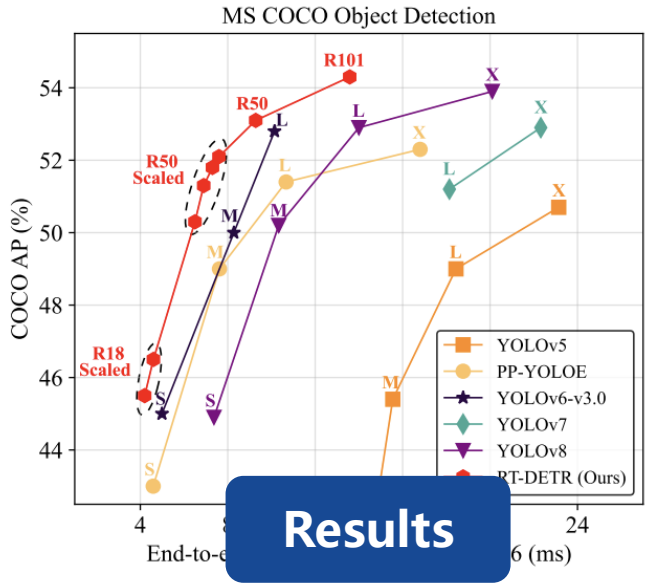
2.2 Real Time DEtection TRansformer

项目代码



Efficient Hybrid Encoder

RT-DETR 的 Efficient Hybrid Encoder 结合 CNN 层级特征提取和Transformer 自注意力机制，通过**跨尺度特征融合和局部注意力计算**，减少全局计算开销，提高目标检测的效率和实时性。



Results

相较于YOLO系列，**RT-DETR-L和X两款检测器均实现了更加出色的性能**，速度上的优势是很明显的，可以说，RT-DETR达到了更好的平衡。尤其是大目标的AP指标，RT-DETR是显著高于YOLO系列的，这也许正是得益于**Transformer的长距离捕捉特征的能力。**

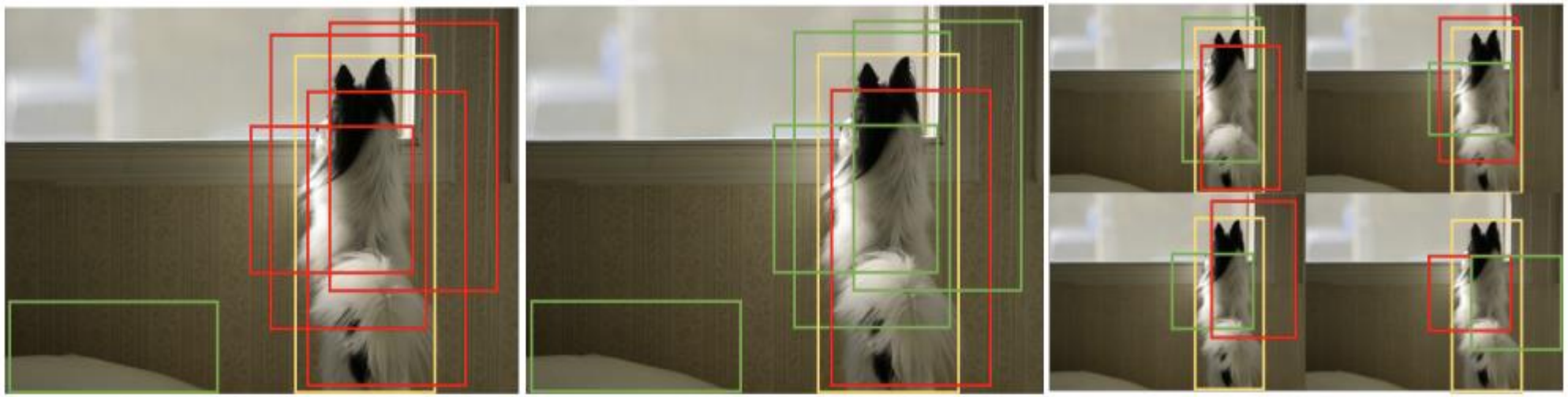
2.3 DEtection Improved Matching

项目代码

提升匹配数量：Dense O2O

背景：在 DETR 及其改进模型中，目标检测依赖 **One-to-One（O2O）**匹配，即通过 匈牙利匹配，让每个预测框唯一匹配一个真实目标。存在**训练困难、召回率低**等问题。

Dense O2O：Dense O2O 允许多个查询匹配同一目标，提高学习效率。**为每个真实目标框分配多个候选查询，计算多个匹配方案**，从而提升训练信号的密集度。



(a) **O2M**: 1 target and 4 pos. (b) **O2O**: 1 target and 1 pos. (c) **Dense O2O by stitching**: 4 targets and 4 pos.

Dense O2O（c）可以提供与 O2M（a）相同质量的正样本。

黄色、红色和绿色框分别代表 GT、正样本和负样本

2.3 DEtection Improved Matching

项目代码

提升匹配质量：Matchability-Aware Loss，MAL

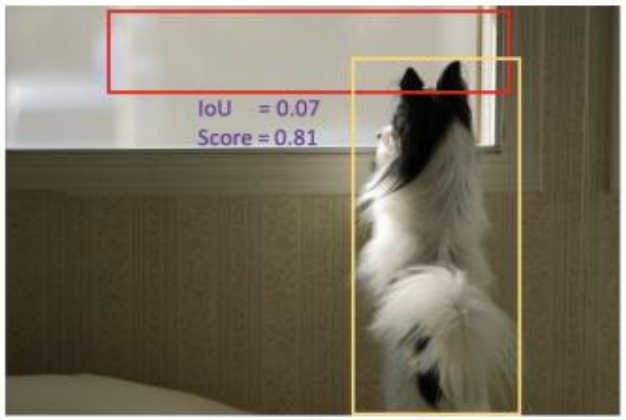
VFL的问题：（1）对于IoU很低的匹配，loss惩罚不会随着置信度增加而增加。（2）IoU=0的时候，这被认为是负样本，进一步减少正样本数量。

MAL改进：MAL在低质量匹配下，会随着置信度越高，惩罚越大。

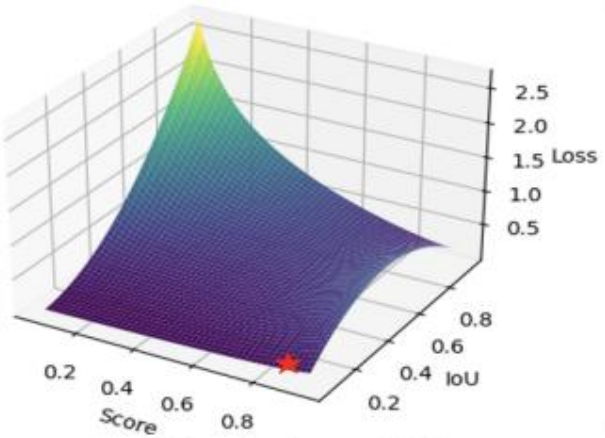
$$\mathrm{VFL}(p,q,y)=\begin{cases}-q(q\log(p)+(1-q)\log(1-p)) & q>0\\ -\alpha p^{\gamma}\log(1-p) & q=0,\end{cases}\tag{3}$$

$$\mathrm{MAL}(p,q,y)=\begin{cases}-q^{\gamma}\log(p)-(1-q^{\gamma})\log(1-p) & y=1\\ -p^{\gamma}\log(1-p) & y=0.\end{cases}\tag{4}$$

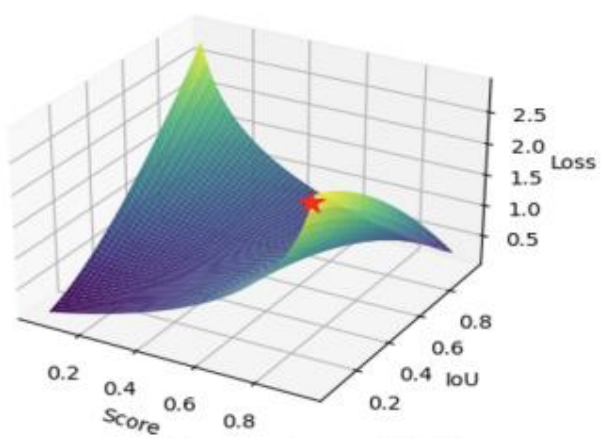
其中 q 表示预测边界框与其目标框之间的 IoU



(d) Low-quality matching



(e) Loss landscape of VFL



(f) Loss landscape of MAL

对于低质量匹配，使用 VFL 和 MAL 时的损失值标记为 ⋆，表明 MAL 可以更有效地优化这些情况v

3

PKU_Campus实验结果

PKU_Campus Experiment Results

3.1 检测结果

项目代码

3.2 实时检测

项目代码

DETR概述

DETR相关论文分析

PKU_Campus实验结果

3.2 定性结果展示

项目代码

DEIM

RT-DETR (Best)

YOLO-v8

3.3 参考文献

1. Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.
2. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
3. Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16965-16974.
4. Huang S, Lu Z, Cun X, et al. DEIM: DETR with Improved Matching for Fast Convergence[J]. arXiv preprint arXiv:2412.04234, 2024.

项目代码

中国科学院大学
University of Chinese Academy of Sciences

汇报结束　　感谢观看

敬请各位老师同学批评指正

汇报人：唐治江

2025年4月5日