doi: 10.7690/bgzdh.2026.05.020

EMOTR：一种高效端到端多目标跟踪框架

傅文鸿[1,2]，曾　刊[1]，王长城[1]，康　林[3]，周子浩[2]
（1. 中国兵器装备集团自动化研究所有限公司系统总体部，四川 绵阳 621000；
2. 国防科技大学智能科学学院，长沙 410073；
3. 陆军装备部驻重庆地区军事代表局驻广元地区军事代表室，四川 广元 628000）

摘要：为解决 Transformer 端到端多目标跟踪（end-to-end multi-object tracking with transformer，MOTR）自注意力计算量随分辨率二次增长，导致训练慢、显存占用高、批量小等问题，提出一种高效端到端多目标跟踪（efficient MOTR，EMOTR）方法。在不改变 MOTR 范式的前提下，快速多尺度注意力（fast multi-scale attention，FMA）以 I/O 友好的注意力内核配合多尺度特征融合，降低计算与存储开销，同时提升小目标分辨；时空批处理与解码器权重共享，实现变长序列同批训练并减少约 15%参数；自动混合精度（automatic mixed precision，AMP）结合动态 Loss Scaling，充分释放 Tensor Core 吞吐。基于 VisDrone2019 的实验结果表明：相较原始 MOTR，训练时间下降 80.4%、参数下降 15.5%，MOTA 提升 2.1 至 24.9 且 IDF_1 保持稳定，验证了在不牺牲精度的前提下显著提升端到端 MOT 实用性的可能性。

关键词：多目标跟踪；Transformer；注意力机制；混合精度；无人机视觉
中图分类号：TP391.41　　**文献标志码：**A

EMOTR: An End-to-end Framework for Efficient and Accurate Multi-object Tracking

Fu Wenhong[1,2], Zeng Kan[1], Wang Changcheng[1], Kang Lin[3], Zhou Zihao[2]
(1. *Department of System General*, *Automation Research Institute Co., Ltd. of China South Industries Group Corporation*, *Mianyang* 621000, *China*; 2. *College of Intelligence Science and Technology*, *National University of Defense Technology*, *Changsha* 410073, *China*; 3. *PLA Military Representative Office in Guangyuan District of Chongqing District Military Representative Bureau of Army Armaments Department*, *Guangyuan* 628000, *China*)

Abstract: To address the issues of slow training, high memory usage, and small batch sizes caused by the quadratic growth of self-attention computation with resolution in Transformer-based end-to-end multi-object tracking (MOTR), an efficient MOTR (EMOTR) method is proposed. Without altering the MOTR paradigm, fast multi-scale attention (FMA) utilizes an I/O-friendly attention kernel combined with multi-scale feature fusion to reduce computational and storage overhead while enhancing small object resolution. Spatio-temporal batch processing and decoder weight sharing enable variable-length sequence training within the same batch and reduce approximately 15% of parameters. Automatic mixed precision (AMP) combined with dynamic Loss Scaling fully utilizes Tensor Core throughput. Experimental results based on VisDrone2019 show that compared to the original MOTR, training time is reduced by 80.4%, parameters are decreased by 15.5%, MOTA is improved by 2.1 to 24.9, and IDF_1 remains stable, verifying the possibility of significantly enhancing the practicality of end-to-end MOT without sacrificing accuracy.

Keywords: multi-object tracking; Transformer; attention; mixed precision; drone vision

0　引言

多目标跟踪（multi-object tracking，MOT）在无人机巡察、态势感知等任务中至关重要。传统“检测-再关联”范式（SORT[1]、DeepSORT[2]、BoT-SORT[3]）检测与关联分离，难以端到端优化。Transformer 端到端多目标跟踪（MOTR）[4]、TrackFormer[5]等验证了 Tracking-by-Query 的优势，但标准注意力计算与显存随 token 数 N 二次增长，高分辨率使训练受限（常为 batch size=1）。现有工作缺少面向无人机多尺度小目标的系统协同设计，为此，笔者提出高效端到端多目标跟踪（EMOTR），协同优化算子级高效注意力、多尺度特征利用、时空批处理与混合精度训练，并通过解码器权重共享增强可训练性。主要贡献：1）提出快速多尺度注意力（FMA），显著降低显存与计算开销并增强小目标表征；2）设计时空批处理与解码器权重共享策略，减少约 15%参数冗余；3）在 VisDrone2019[6]上验证训练时间降低 80.4%，MOTA 提升 2.1 至 24.9，且 IDF_1

收稿日期：2024-12-11；**修回日期：**2025-01-25
第一作者：傅文鸿（1999—），男，福建人，硕士。

保持稳定。

1 EMOTR 高效端到端多目标跟踪方法

给定长度 T 的视频序列，每帧输出目标集合及身份标识。沿用 MOTR 的 Tracking-by-Query 设定，主干网络在不同尺度 l 产生特征图，编码器全局建模，解码器以查询输出分类与边界框。令多尺度 token 数 $N=\sum_l H_l W_l$，特征维度 d，标准注意力复杂度 $O(N^2d)$，显存 $O(N^2)$，是高分辨率训练的主要瓶颈。如图 1 所示，EMOTR 整体框架：主干提取多尺度特征经编码器获得全局上下文；解码器接收上一帧跟踪查询与检测查询，通过交叉注意力完成定位与身份更新；轨迹增删模块根据分数阈值完成轨迹生成、更新与终止。该分模块设计使得笔者提出的优化方法均可通过消融实验独立验证。

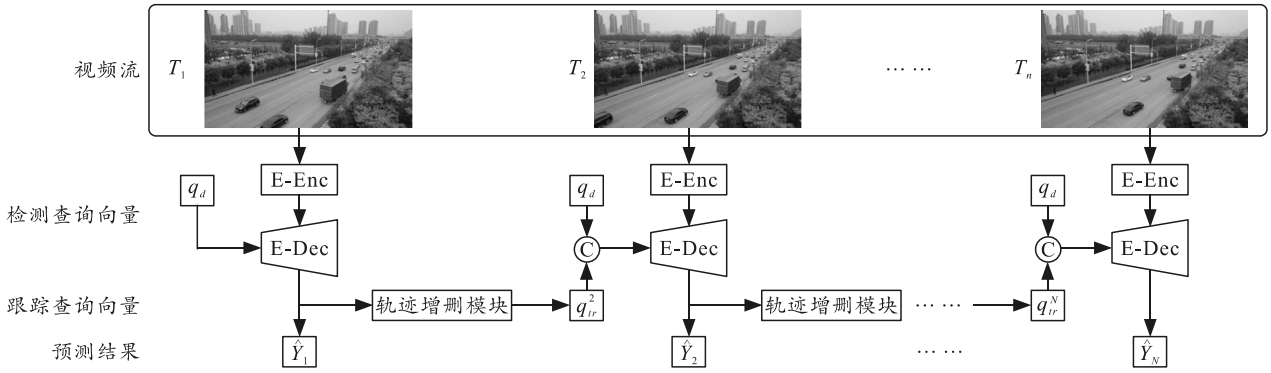

图 1 EMOTR 的总体框架

1.1 FMA

解码器注意力采用缩放点积注意力：

$$\mathrm{Attn}(Q,K,V)=\mathrm{Softmax}\left(QK^{\mathrm{T}}/\sqrt{d}\right)V。\quad(1)$$

式中：Q、K、V 分别为查询、键和值；d 为特征维度。FMA 采用 FlashAttention[7]作为高效计算内核，通过分块计算与片上 SRAM 重用避免完整注意力矩阵写回 HBM，获得 1～2×算子级加速。在此基础上进行多尺度特征集成，对每帧提取 $C_3/C_4/C_5$ 层特征映射到统一维度后作为键/值集合：

$$\mathrm{Attn}_m s(Q,K^l,V^l)=\Sigma_l\mathrm{Softmax}\left(Q(K^l)^{\mathrm{T}}/\sqrt{d}\right)V^l。\quad(2)$$

式中 l 为尺度层级。远距离小目标由高分辨率层提供定位线索，近距离与遮挡目标由高语义层提供稳定表征，减少额外特征金字塔后处理。

1.2 时空批处理与解码器权重共享

提出时空批处理策略：对若干段视频片段按最长 $T_{\max}$ 进行 padding，构造时间掩码：

$$M_{i,t}=1(t\leqslant T_i)，否则\ M_{i,t}=0。\quad(3)$$

训练时将掩码融入注意力 mask，使填充帧不参与 Softmax 归一化，实现变长序列同批训练。同时在时间维对解码器采用循环权重共享，将 L 层独立参数约束为共享同一组，减少约 15%参数并起正则化作用。

1.3 混合精度训练与稳定性

引入自动混合精度(AMP)，将矩阵乘法、卷积等算子置于 FP16 路径，参数更新与归一化保持 FP32。采用动态 Loss Scaling 保证收敛稳定性。

1.4 损失函数

总损失函数：

$$\mathcal{L}=\lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}}+\lambda_{L_1}\mathcal{L}_{L_1}+\lambda_{\mathrm{GIoU}}\mathcal{L}_{\mathrm{GIoU}}。\quad(4)$$

式中 $\mathcal{L}_{\mathrm{cls}}$ 采用 Focal Loss[8]缓解正负样本不平衡，回归由 L_1 与 GIoU[9]组成。训练采用匈牙利匹配建立预测与真实目标一一对应。模型含轨迹查询 q_{tr} 与检测查询 q_d，前者携带身份信息实现跨帧递归更新，后者捕获新目标。轨迹增删模块根据置信度阈值进行轨迹生成、更新与终止。

2 实验与数据分析

2.1 数据集与指标

在无人机基准 VisDrone2019[6]上验证，采用 CLEAR MOT 指标[10]：

$$\mathrm{MOTA}=1-(\mathrm{FN}+\mathrm{FP}+\mathrm{IDS})/\mathrm{GT};\quad(5)$$

$$\mathrm{IDF_1}=2\bullet\mathrm{IDTP}/(2\bullet\mathrm{IDTP}+\mathrm{IDFP}+\mathrm{IDFN})。\quad(6)$$

式中：GT 为真实目标总数；IDS 为统计身份切换次数，衡量轨迹身份稳定性；$\mathrm{IDF_1}$ 为衡量身份一致性的匹配质量；IDTP、IDFP、IDFN 分别为身份匹配意义下的真阳性、假阳性、假阴性。

2.2 实验设置

所有实验基于同一代码库，训练与推理在单张 NVIDIA RTX 3090 上进行。优化器 AdamW，学习

率余弦退火策略与基线保持一致。

2.3　消融实验

消融结果如表 1 所示。

表 1　EMOTR 消融结果（VisDrone2019 验证集）

模型	MOTA	IDF_1	参数量/M	IDS	训练时间/(min/epoch)
A：基线 MOTR	22.8	41.4	45.1	959	180.3
B：*A*+批处理&共享	22.7	41.1	37.5	964	96.2
C：*B*+AMP	22.7	41.1	37.5	1 796	48.3
D：*C*+(FMA)	24.9	41.5	38.1	1 823	35.3

表 1 给出逐步引入各模块的消融结果。*A* 为原始 MOTR 基线，训练时间 180.3 min/epoch。*B* 引入时空批处理与权重共享后降至 96.2 min/epoch（约降 46.7%），参数由 45.1 降至 37.5 M。*C* 启用 AMP 后降至 48.3 min/epoch（相对 *B* 再降约 49.8%），采用动态 Loss Scaling 保证稳定性。*D* 加入 FMA 后降至 35.3 min/epoch（总计降 80.4%），MOTA 从 22.8 提升到 24.9，表明多尺度特征集成改善了小目标召回。*D* 的 IDS（1 823）相比 *A*（959）有所上升，主因是更高召回导致目标频繁进出视野，使轨迹重启概率增加，但 IDF_1 保持稳定说明身份匹配质量未退化。综上，EMOTR 训练时间下降 80.4%、参数下降 15.5%，MOTA 提升至 24.9，验证了各模块协同优化的有效性。

2.4　与现有方法比较

EMOTR 与现有方法比较如表 2 所示。

表 2　EMOTR 与现有方法比较

方法	MOTA	IDF_1	IDS
MOTDT	−0.8	21.6	1 437
SORT	14.0	38.0	3 629
YOLOv9[11]+BoT−SORT	19.8	38.5	568
Visual−Spatial	20.7	32.5	635
UTOPIA	22.8	37.5	503
MOTR（基线）	22.8	41.4	959
TrackFormer	24.5	30.5	4 840
EMOTR（本文中方法）	24.9	41.5	1 823

EMOTR 在 MOTA 与 IDF_1 上取得最优。MOTA 提升主要来自 FMA 多尺度集成降低漏检。IDS 上升系召回提升后目标频繁进出视野所致，IDF_1 稳定说明身份质量未退化。在巡察与搜索等应用中，EMOTR 在精度与效率间取得了更适合工程部署的平衡。

2.5　复杂度与显存分析

标准注意力需构造 $N\times N$ 矩阵，计算约 $O(N^2d)$，显存约 $O(N^2)$。FMA 使用 FlashAttention 以分块方式在 SRAM 完成 Softmax 与加权求和，缓解显存峰值。FMA 结合权重共享与更大批量，使算子级与数据并行优化叠加，整体训练时间减少约 80%。

3　结论

笔者提出 EMOTR，通过 FMA、时空批处理与 AMP 3 项协同优化，在不改变 MOTR 端到端范式前提下显著提升可训练性与精度。在 VisDrone2019 上，训练时间降低 80.4%、参数降低 15.5%，MOTA 提升 2.1 至 24.9 且 IDF_1 保持稳定，验证了 Transformer 式 MOT 框架的工程可行性。EMOTR 可广泛应用于边境巡逻与安防监控、灾害搜救与应急响应、交通流量监测以及智慧城市管理等场景，其高效训练特性使模型能快速适配不同任务需求，具有显著的应用创新性与工程推广价值。

参考文献：

[1] BEWLEY A, GE Z, OTT L, et al. Simple online and realtime tracking[C]// 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016: 3464−3468.

[2] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric[C]// 2017 IEEE International Conference on Image Processing (ICIP). Beijing: IEEE, 2017: 3645−3649.

[3] AHARON N, ORFAIG R, ZELNIK-MANOR L. BoT-SORT: Robust associations multi-pedestrian tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2022: 1−18.

[4] ZENG F G, DONG B, ZHANG Y, et al. MOTR: End-to-end multiple-object tracking with transformer[C]// European Conference on Computer Vision (ECCV 2022). Cham: Springer, 2022: 659−675.

[5] MEINHARDT T, KIRILLOV A, LEAL-TAIXÉ L, et al. TrackFormer: Multi-object tracking with transformers[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 8844−8854.

[6] ZHU P F, WEN L W, BIAN X, et al. Vision meets drones: A challenge[C]// Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018: 1−24.

[7] DAO T, FU D Y, ERMON S, et al. FlashAttention: Fast and memory-efficient exact attention with IO-awareness[C]//NeurIPS. 2022: 16223−16236.

[8] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 2980−2988.

[9] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE,

2019: 658-666.
[10] BERNARDIN K, STIEFELHAGEN R. Evaluating multiple object tracking performance: the CLEAR MOT metrics[J]. EURASIP, 2008, 2008(1): 1-10.
[11] WANG C Y, YEH I H, LIAO H Y M. YOLOv9: Learning what you want to learn using programmable gradient information[C]// European Conference on Computer Vision (ECCV 2024). Cham: Springer, 2024: 1-21.

（上接第 74 页）

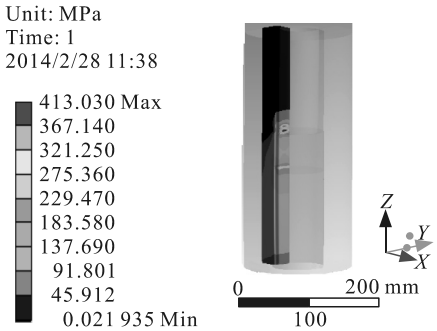

图 6 成型块等效应力

由图 6 可得，成型块最大等效应力 σ_{max}=413.03 MPa＜σ_b。剩余强度系数 $\eta=\sigma_b/\sigma_{max}$=3.10。

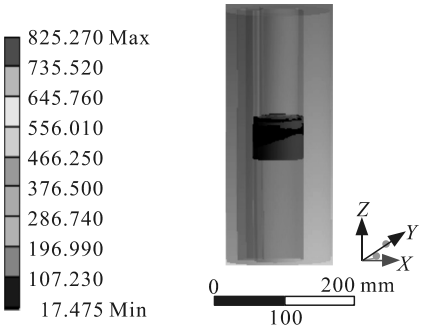

图 7 压块等效应力

由图 7 可得，最大等效应力发生在压块上，σ_{max}=825.27 MPa＜σ_b。剩余强度系数 $\eta=\sigma_b/\sigma_{max}$=1.55。

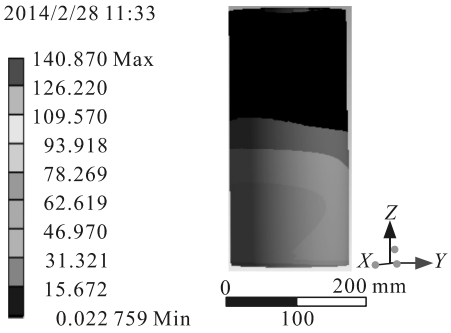

图 8 内套等效应力

由图 8 可得，内套最大等效应力 σ_{max}=140.87 MPa＜530 Mpa。剩余强度系数 $\eta=\sigma_b/\sigma_{max}$=3.76。

3.3 仿真结果

根据以上仿真计算及分析，该种异形药柱压药模具各零部件强度满足使用要求，可以据此开展模具图纸设计及加工。

4 模具验证

通过理论计算及仿真计算，该种结构的异形药柱模具强度满足使用要求。根据图纸加工了模具，为了验证模具的强度情况，采用惰性材料进行模拟压制，压制后模具完好无损伤；模具经惰性材料验证后，使用炸药进行药柱压制，压制后模具完好无损伤，成型药柱的质量、密度及尺寸满足设计要求，提升了异形药柱压制过程的安全性。

5 结论

在加工异形模具前，对模具进行了理论计算及仿真计算，计算结果满足模具安全性要求；加工模具后，采用惰性材料和炸药进行了模具强度验证。结果表明：模具强度满足要求，理论计算及仿真计算符合实际情况。该研究已在某产品异形药柱压制中成功应用，该产品异形药柱压制过程安全，可以在其他异形药柱压制过程中推广应用。

参考文献：

[1] 王泽山，张丽华，杨春海. 装药工程[M]. 北京：北京理工大学出版社, 2010: 80-110.
[2] 陈国光，董素荣. 弹药制造工艺学[M]. 北京：北京理工大学出版社, 2004: 331-345.
[3] 肖忠良，胡双启，吴晓青，等. 火炸药的安全与环保技术[M]. 北京：北京理工大学出版社, 2006: 1-11.
[4] 《飞机设计手册》总编委会. 飞机设计手册：第 9 册载荷、强度和刚度[M]. 北京：航空工业出版社，2001: 215-264.
[5] 赵晓梅，闫光虎，严文荣，等. ANSYS 在发射药力学性能仿真模拟中的应用[J]. 兵工自动化，2012，31(7): 35-38.