GTELT CVPR'22
Motivation

Global Tracking via Ensemble of Local Trackers

Zikun Zhou[1,*], Jianqiu Chen[1,*], Wenjie Pei[1,†], Kaige Mao[1], Hongpeng Wang[1,2], and Zhenyu He[1,†]
[1]Harbin Institute of Technology, Shenzhen [2]Peng Cheng Laboratory

长时跟踪的一大难点: 目标的非连续运动（遮挡、消失）

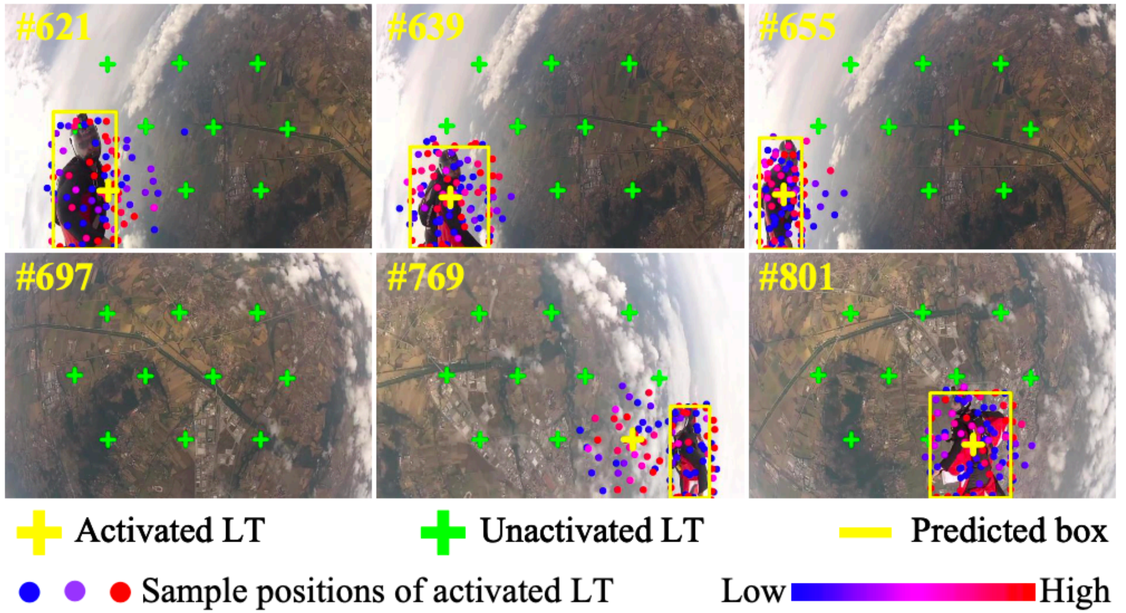
(1) 短时跟踪 + 全局重检测 (LTMU)

- 当目标消失时，短时跟踪器易将干扰物误判做目标，而不是启动重检测器

(2) 每帧做全局检测 (GlobalTrack)

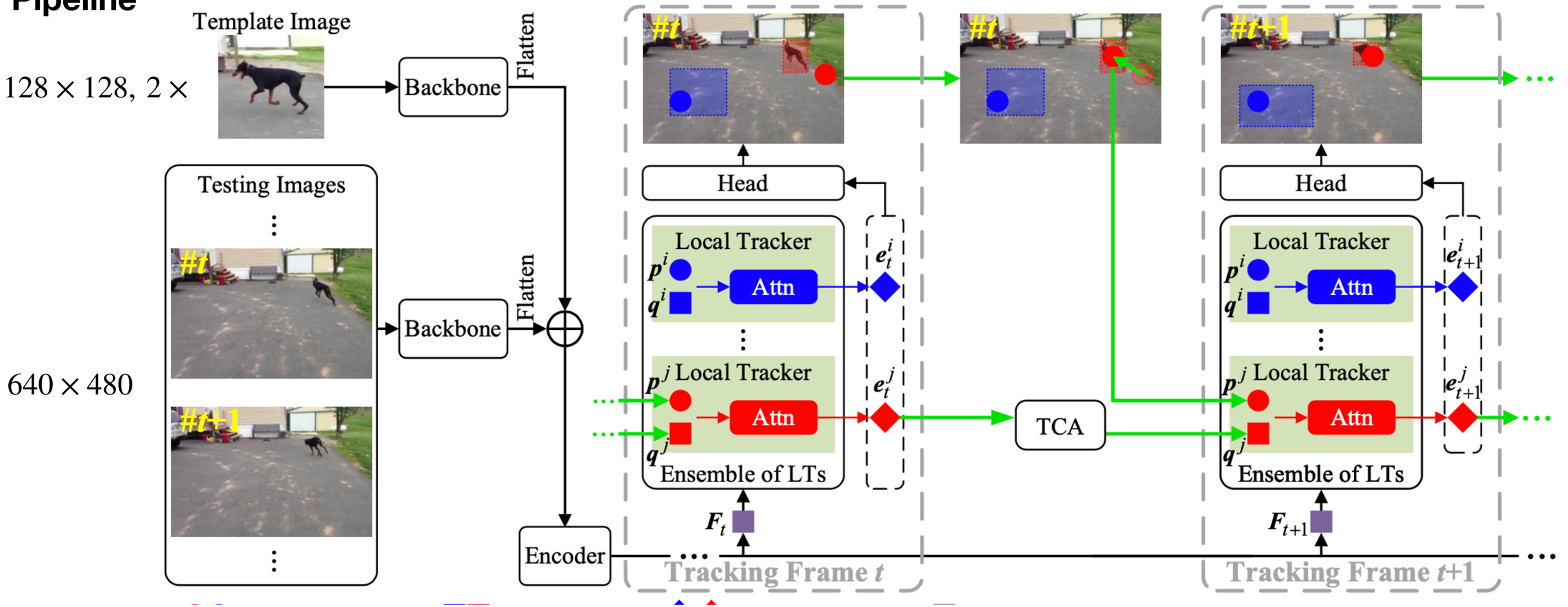
- 没有利用时序信息，易受表观变化影响

本方法：基于Deformable DETR，将 N 个 object query 与 reference point 视作在全图散布的 N 个局部跟踪器。一个局部跟踪器持续跟踪目标，跟踪失败时，启动其它局部跟踪器。



GTELT CVPR'22

Pipeline



1. 每个局部跟踪器保存一个 **target query** q^i (表观信息) 和一个 **reference point** p^i (位置信息)
2. target query 之间进行 **self-attention**，使 reference points 尽可能在全图铺开
3. 根据 target query 预测相对于 reference point 的 offset，再与 target query 进行 **deformable cross-attention**

$$\{\boldsymbol{e}_t^i\}_{i=1}^N = \Phi_{\mathrm{LT}}(\{\boldsymbol{q}^i\}_{i=1}^N, \{\boldsymbol{p}^i\}_{i=1}^N\}, \boldsymbol{F}_t),$$

$$\{y_t^i\}_{i=1}^N = \Phi_{\mathrm{Head}}(\{\boldsymbol{e}_t^i\}_{i=1}^N).$$

$$\boldsymbol{p}^j \leftarrow \mathcal{T}_{\mathrm{rp}}(y_t^j), \boldsymbol{q}^j \leftarrow \mathcal{T}_{\mathrm{tq}}(\boldsymbol{e}_t^j),$$

GTELT CVPR'22

Temporal Context Transfer

1. reference point: 前一帧的跟踪结果 (cx, cy) 作为下一帧被激活跟踪器的 reference point；当跟踪失败时，重置被激活跟踪器的 reference point
2. target query: Temporal Context Aggregation (TCA) 模块，利用 cross-attention 融合历史信息

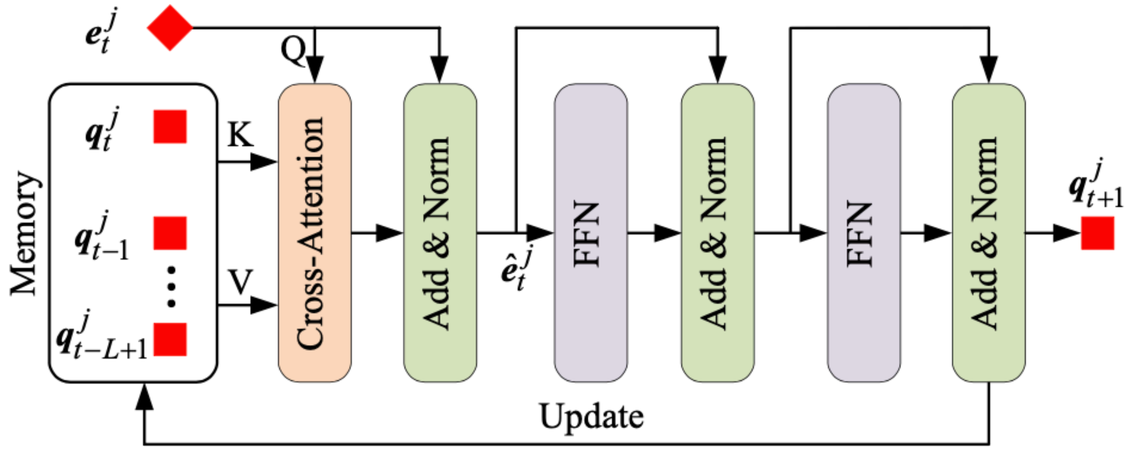


Figure 3. **Architecture of the proposed temporal context aggregation model.** It performs the interaction between the target embedding and the historical online target queries to aggregate the temporal context information modeled by these queries.

Training

1. 类似 DETR，利用匈牙利算法匹配 GT 与预测结果
2. reference point 与 GT center 之间的 L1 Norm，鼓励同一个局部跟踪器持续跟踪目标，deformable cross-attention 的范围合理
3. 所有结果求分类损失，匹配结果额外求回归损失

$$\mathcal{L}_H(y_t^i, \hat{y}_t) = \lambda_{cls}\mathcal{L}_{cls}(s_t^i) + \mathcal{L}_{box}(\boldsymbol{b}_t^i, \hat{\boldsymbol{b}}_t) + \lambda_r\mathcal{L}_{\ell_1}(\boldsymbol{p}^i, \hat{\boldsymbol{p}}_t),$$

$$\mathcal{L}_{box}(\boldsymbol{b}_t^i, \hat{\boldsymbol{b}}_t) = \lambda_{\ell_1}\mathcal{L}_1(\boldsymbol{b}_t^i, \hat{\boldsymbol{b}}_t) + \lambda_{iou}\mathcal{L}_{iou}(\boldsymbol{b}_t^i, \hat{\boldsymbol{b}}_t).$$

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\lambda_{cls}\mathcal{L}_{cls}(s_t^i) + \mathbb{1}_{\{i=\pi_t\}}\mathcal{L}_{box}(\boldsymbol{b}_t, \hat{\boldsymbol{b}}_t^i),$$

GTELT CVPR'22
Experiments

对比实验 (GTELT / Stark-101 / OSTrack-384):

- LaSOT: 67.7 / 67.1 / **71.1**
- LaSOT_ext: 45.0 / - / **50.5**
- TrackingNet: 82.5 / 82.0 / **83.9**
- FPS: 26.0 (2080Ti) / 31.7 (V100) / **58.1 (2080Ti)**

Table 2. **Ablation studies on the local tracker number (N) and the memory length (L) on LaSOT.**

	L=5			N=10				
	N=5	N=10	N=20	L=1	L=3	L=5	L=7	L=9
Pre.	0.712	0.732	0.705	0.725	0.726	0.732	0.731	0.730
nPre.	0.743	0.759	0.734	0.755	0.755	0.759	0.759	0.757
AUC	0.670	0.677	0.660	0.672	0.675	0.677	0.677	0.675

Table 1. **Precision (Pre.), normalized precision (nPre.), and AUC for four variants of our method on LaSOT.**

Variants	Our model	OSDet	EGT	SGT
Pre.	0.732	0.707	0.690	0.670
nPre.	0.759	0.732	0.717	0.705
AUC	0.677	0.653	0.623	0.619

1) **Our model**, our intact model that performs global tracking via ensemble of local trackers.
2) **OSDet**, which removes the temporal context transferring scheme from our model. Thus, our model degenerates into a global **O**ne-**S**hot **Det**ector in the transformer framework, using multiple offline learned queries to perform global redetection without exploiting the temporal context.
3) **EGT**, replacing the deformable attention module in our model with the multi-head attention module, which computes attention globally and densely. Thus, the ensemble of local trackers in our model becomes an **E**nsemble of **G**lobal **T**rackers. Due to these global trackers do not involve reference positions, we perform temporal context transferring only using the target queries as the carriers.
4) **SGT**, reducing the number of target queries in EGT to one, i.e., adopting a **S**ingle multi-head attention-based **G**lobal **T**racker to perform tracking.