《基于混合注意力的 Transformer 视觉目标跟踪算法》附录

附录 A　实验相关内容

1. 为更好利用时间线索并适应目标外观的变化，对模板进行动态更新。具体地，初始化时设置模板集合大小为 n，跟踪时将第一帧输入至模板分支，并对其进行数据扩充策略，构建包含 n 个样本的集合，用于初始跟踪。在跟踪过程中，每 5 帧将根据置信度分数保存的模板特征添加到集合中，并将集合中最旧的模板删除。一旦模板集合更新，将该集合$T \in R^{n\times C\times H\times W}$作为 Template Feature 输入至 Transformer 编码器中，通过编码器计算新的聚合特征 Encoded Template Feature，来相互增强多个模板特征，这些高质量的编码特征有利于跟踪模型的生成，然后将该编码聚合特征传播到解码器中，与当前的搜索特征进行交叉注意力计算，在跟踪过程中每一帧都使用到解码器，来不断产生 Decoded Search Feature 用于后续的回归与分类，由此来实现模板帧和搜索帧的交互，使目标的搜索更精确。
2. 其中 SiamRPN 将目标检测领域的区域推荐网络引入跟踪，ATOM 采用 IoUNet 提出重叠区域最大化的训练方式，DiMP50 以端到端的方式增强了学习到的 CNN 内核的判别能力，PrDiMP50 提出基于概率的回归方法并将其用于跟踪。

   1) 背景杂乱。跟踪过程中目标与背景的可区分程度直接影响到跟踪器的性能，如视频序列 Basketball，Coupon。其中 Coupon 序列从 198 帧起仅本文算法能够完全正确跟踪到目标。当背景包含与目标相近的影响物时，大多数跟踪器会发生漂移使跟踪失败，而本文算法通过注意力机制增强对特征的表达能力，以及对局部与全局信息的捕捉，使跟踪器更好利用目标的特征，仍能较好跟踪目标。

   2) 目标遮挡。目标遮挡是跟踪中常见的问题。如视频序列 Bolt 和 Coupon，目标在运动过程中会出现部分遮挡，这时模型比较容易漂移,造成跟踪失败。本文通过 Transformer 的编解码器分支对多个模板帧输入和搜索帧进行处理，加强了对模板的利用，使跟踪器更为鲁棒，有效减小遮挡物带来的影响。

   3) 尺度变化。以视频序列 Coupon 和 Diving 为例，目标在跟踪过程中均发生了明显的尺度变化。Diving 序列在 99 帧时目标发生巨大形变，PrDiMP50、SiamRPN、DiMP50 三个跟踪器皆无法准确捕捉到目标，包围框包含大量背景信息，而相比于 ATOM，本文算法跟踪更为准确，这证明了本文算法对于目标尺度变化有较好的应对能力。

   4) 平面外旋转。视频序列 Basketball、Bolt、Diving 均有该属性。如 Basketball 序列在 629 帧前发生平面外旋转，除本文算法外，其余四个算法均跟踪失败。Bolt 序列在 136 帧发生平面外旋转，除本文算法与 SiamRPN 算法能够正常跟踪目标外，其余算法都出现跟踪框不准确的情况。这得益于本文有效捕捉了目标的全局依赖关系，可以精确定位目标并且避免了背景的影响。

附录 B　相关图表

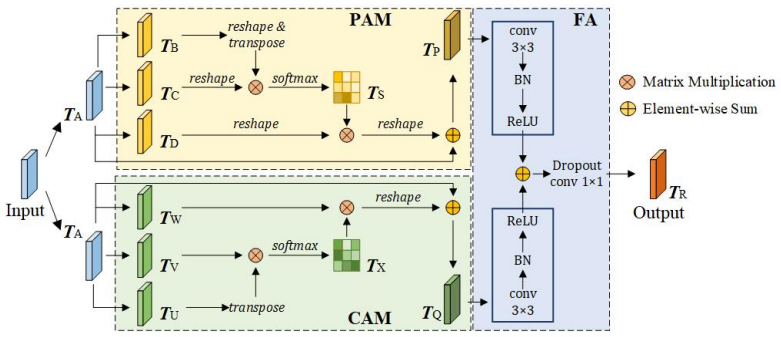

图 B1　混合注意力模块结构图

表 B1　GOT-10k 数据集跟踪结果

	CCOT	ECO	SPM	SiamFC	ATOM	DiMP50	PrDiMP50	TrSiam	Ours
$SR_{0.5}$(%)	32.8	30.9	59.3	35.3	63.4	71.7	73.8	76.6	76.8
$SR_{0.75}$(%)	10.7	11.1	35.9	9.8	40.2	49.2	54.3	57.1	57.4
AO(%)	32.5	31.6	51.3	34.8	55.6	61.1	63.4	66.0	66.7

表 B2　TrackingNet 数据集跟踪结果

	SiamFC	UPDT	D3S	ATOM	DiMP50	SiamRPN++	PrDiMP50	TrDiMP	Ours
Precision (%)	53.3	55.7	66.4	64.8	68.7	69.4	70.4	73.1	74.0
N. Prec. (%)	66.6	70.2	76.8	77.1	80.1	80.0	81.6	83.3	83.9
Success (%)	57.1	61.1	72.8	70.3	74.0	73.3	75.8	78.4	78.8