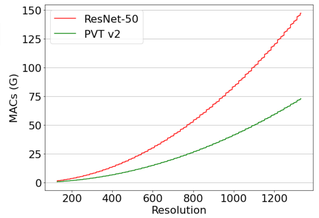
Efficient Joint Detection and Multiple Object Tracking with Spatially Aware Transformer

Siddharth S. Nijhawan, Leo Hoshikawa, Atsushi Irie, Masakazu Yoshimura, Junji Otsuka, Takeshi Ohashi (Sony Group Corporation)

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

- Recent works like TransTrack [1] achieve state-of-the-art MOT performance, however, **struggle to solve the computational bottleneck**, resorting to **HW inefficient operations and modules**.
- **Layer-by-layer profiling** of TransTrack [1] reveals 2 key bottlenecks requiring **78% total parameters** with **92% MACs**: CNN backbone and the Transformer Encoder



- TransTrack [1] uses a **CNN backbone** to extract feature maps but **scales poorly** with increase in the input resolution. Use of a **fully-transformer** based architecture leads to a significant reduction in model size and complexity.
- We **replace the encoder FFN** with a proposed block based on depth-wise convolution operation **reducing its MACs by 86.76%**.

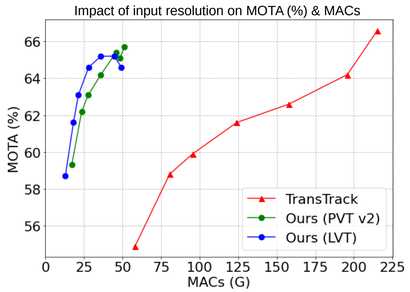
Results

Model	MOTA↑	FP↓	FN↓	IDS↓	Params↓	MACs↓
TransTrack [10]	74.50	28,323	112,137	3,663	46.87M (-0%)	215.23G (-0%)
GTR [21]	75.30	36,231	93,150	**2,346**	43.80M	-[1]
CenterTrack [20]	67.80	**18,489**	160,332	3,039	19.32M	70.88G
FairMOT [19][2]	73.70	27,507	117,477	3,303	19.71M	84.98G
TransCenter [17]	**76.20**	40.101	**88.827**	5.394	35.1M	199.59G
TransCenter-Lite [17]	73.50	-	-	-	8.1M	80.53G
Ours (PVT v2 [14])	73.20	28,341	118,689	4,218	19.34M (-58.73%)	45.80G (-78.72%)
Ours (LVT [18])	71.00	32,730	125,274	5,757	**6.07M (-87.04%)**	**28.82G (-86.60%)**

Impact of Backbone* on Parameters & MACs
(MOT17 half-half set)

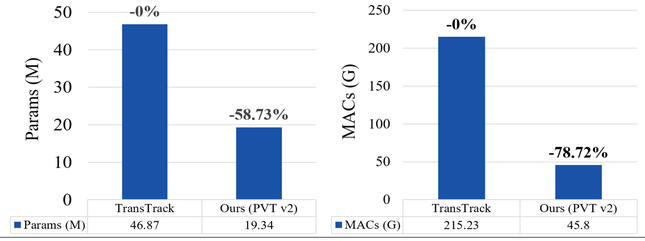
Backbone	MOTA↑	Params↓	MACs↓
ResNet-50 [4]	**66.30**	30.12M	80.04G
PVT v2 [14]	65.70	19.34M	45.80G
LVT [18]	64.60	**6.07M**	**28.82G**

*includes proposed modifications in the architecture

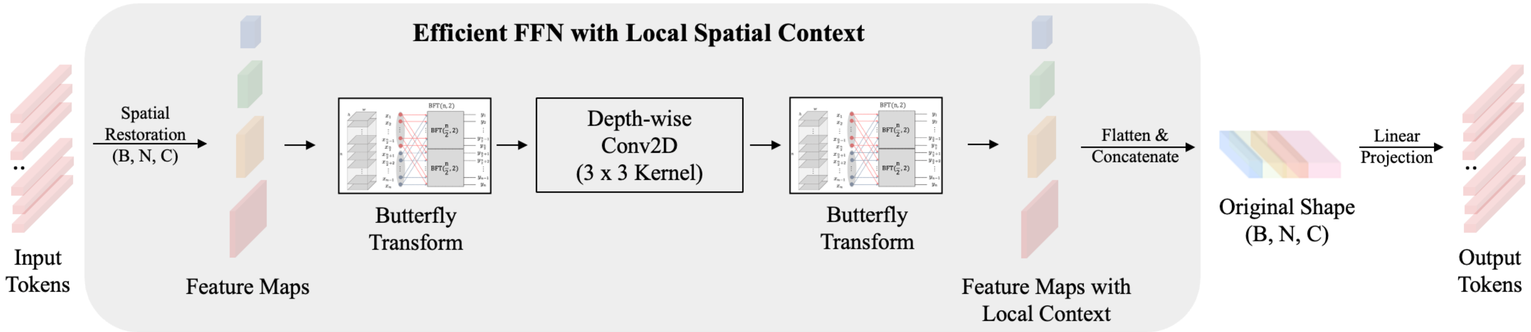


Summary / Conclusion

- We propose a **fully-transformer based Joint Detection and Tracking MOT pipeline** which is a light-weight and highly efficient version of TransTrack [1] optimizing 2 key bottlenecks in its architecture: **CNN Backbone** and the **Transformer Encoder**.
- We also propose a novel drop-in replacement of FFN in transformer encoder which performs **channel fusion with logarithmic complexity** (instead of quadratic) using **Butterfly Transform** [3], and **learns the spatial context** within feature maps, otherwise not present in the multi-head self-attention layer.
- Our architecture contains **58.73% less** parameters and requires **78.72% less MACs** in comparison to TransTrack [1], while achieving state-of-the-art **MOTA of 73.20%**.



Proposed drop-in replacement of FFN in Transformer Encoder



References

[1] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer, 2020.

[2] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with pyramid vision transformer. Computational Visual Media, 8(3):415–424, mar 2022.

[3] Keivan Alizadeh Vahid, Anish Prabhu, Ali Farhadi, and Mohammad Rastegari. Butterfly transform: An efficient fft based neural architecture design, 2019.

[1]Due to incompatibilities in the implementation, we could not compute the MACs for GTR
[2]FairMOT does not include post-processing computation