Track Targets by Dense Spatio-Temporal Position Encoding

Jinkun Cao[1] Hao Wu[2] Kris Kitani[1]

[1]Carnegie Mellon University [2]ByteDance

BMVC 2022

The 33rd British Machine Vision Conference
21st - 24th November 2022, London, UK

Motivation and Goal

1-D positional encoding vectors lose spatial information in vision tasks. We thus propose a new scheme of positional encoding to encode object spatio-temporal position information for multi-object tracking.

Dense Spatio-Temporal (DST) Position Encoding

- Encode object position in a pixel-wise dense fashion.
- Extend from single-frame position to a trajectory of an object.

Three levels of DST position encoding

Image level: given the feature maps of an image ($C \times H \times W$):

$$P(x,y,i) = \begin{cases} -cos\left[\left(\frac{x}{W}+\frac{y}{WH}\right)\pi+\frac{2i\pi}{C}\right], i=2k+1 \\ cos\left[\left(\frac{y}{H}+\frac{x}{WH}\right)\pi+\frac{2i\pi}{C}\right], i=2k \end{cases}$$

Object level: in RoI of objects ($C \times H_R \times W_R$):

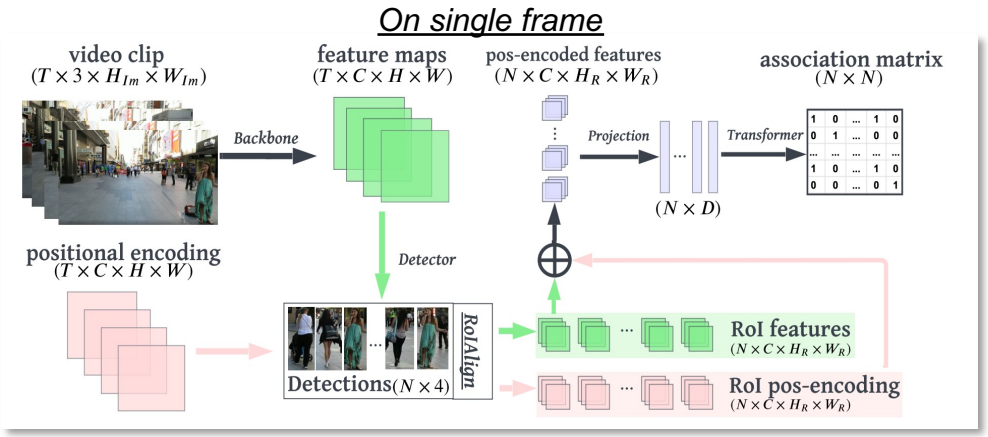
$$P_R(x',y',i) = \begin{cases} -cos\left[\left(\frac{w}{WW_R}x'+\frac{h}{WHH_R}y'\right)\pi+\left(\frac{u}{W}+\frac{v}{WH}\right)\pi+\frac{2i\pi}{C}\right], i=2k+1 \\ cos\left[\left(\frac{h}{HH_R}y'+\frac{w}{WHW_R}x'\right)\pi+\left(\frac{v}{H}+\frac{u}{WH}\right)\pi+\frac{2i\pi}{C}\right], i=2k \end{cases}$$

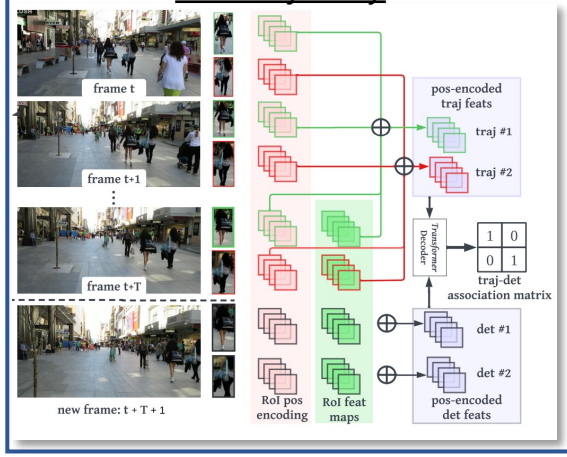
Trajectory level: for an object trajectory $(b_1, b_2, \ldots, b_T)$:

$$P_R^{\mathbf{b}_T|\ldots|\mathbf{b}_1}(x',y',i) = \sum_{t=1}^{T} \alpha_t P_R^{\mathbf{b}_t}(x',y',i),$$

Architecture

The model architecture uses an off-the-shelf detector to gain object positions and a transformer to compare object features for association in tracking across multiple frames.





Experiments

Table 2: Results on MOT17 test set. Best results among transformer methods are underlined.

Tracker	Transformer	HOTA ↑	AssA ↑	MOTA↑	IDF1 ↑	ID Sw. ↓	FP ↓	FN ↓
FairMOT [44]		59.3	58.0	73.7	72.3	3,303	27,507	117,477
PermaTrack [28]		55.5	53.1	73.8	68.9	3,699	28,998	115,104
TraDes [29]		52.7	50.8	69.1	63.9	3,555	20,892	150,060
TubeTK [20]		48.0	45.1	63.0	58.6	4,137	27,060	177,483
ByteTrack [43]		63.1	62.0	**80.3**	77.3	2,196	25,491	**83,721**
OC-SORT [8]		**63.2**	**63.4**	78.0	**77.5**	**1,950**	**15,129**	107,055
TransTrk[30]	✓	54.1	47.9	75.2	63.5	4,614	50,157	86,442
TransCenter [33]	✓	54.5	49.7	73.2	62.2	3,663	23,112	123,738
TrackFormer [23]	✓	-	-	65.0	63.9	3,258	70,443	123,552
MOTR [36]	✓	-	-	67.4	67.0	1,992	32,355	149,400
GTR [40]	✓	59.1	61.6	75.3	71.5	2,859	26,793	109,854
MeMOT [8]	✓	56.9	55.2	72.5	69.0	2,724	37,221	115,248
Ours	✓	60.1	62.1	75.2	72.3	2,729	24,227	109,912

Table 3: Results on DanceTrack test set. Best transformer-based results are underlined.

Tracker	Transformer	HOTA ↑	DetA ↑	AssA ↑	MOTA↑	IDF1 ↑
CenterTrack [44]		41.8	78.1	22.6	86.8	35.7
FairMOT [44]		39.7	66.7	23.8	82.2	40.8
SORT [8] + YOLOX [8]		47.9	72.0	31.2	**91.8**	50.8
DeepSORT [43] + YOLOX [8]		45.6	71.0	29.7	87.8	47.9
ByteTrack [43] + YOLOX [8]		47.3	71.6	31.4	89.5	52.5
OC-SORT [8] + YOLOX [8]		**55.1**	**80.3**	**38.0**	89.4	**54.2**
TransTrk[30]	✓	45.5	75.9	27.5	88.4	45.2
MOTR [36]	✓	48.4	71.8	32.7	79.2	46.1
GTR [40]	✓	48.0	72.5	31.9	84.7	50.3
Ours	✓	51.9	72.3	34.6	84.9	51.0

Table 4: The ablation study of **positional encoding** on MOTS20-val.

pos-encode	HOTA ↑	IDF1 ↑	DetA ↑	AssA ↑	sMOTA↑	MOTSA ↑	ID Sw.↓
w/o pos-encoding	64.4	72.5	72.5	58.0	71.6	82.8	150
classic pos-encoding [24]	64.1	72.5	69.7	59.3	67.8	79.6	162
DST pos-encoding	67.1	74.9	72.8	62.3	71.7	83.0	135

Conclusion

we propose a novel **Dense Spatio-Temporal (DST) position encoding** to incorporate object position information and appearance into the transformer for multi-object tracking. While multiple previous works have failed in boosting performance with classic positional encoding, our work provides a novel and effective paradigm for future works.