Online Multiple Object Tracking with Cross-Task Synergy

Song Guo*¹, Jingya Wang*¹,², Xinchao Wang³,⁴, Dacheng Tao¹

¹The University of Sydney, ²ShanghaiTech University,
³National University of Singapore, ⁴Stevens Institute of Technology
Tracking-by-Detection

Frame t

Association

Frame t + 1
Motivation

**Existing Methods**: The two tasks do not benefit each other
Main Idea

Our Method: Bridge the two tasks to bring synergy
Model

(a) Position Prediction by Regression
(b) Temporal-Aware Target Attention & Distractor Attention
(c) Identity-Aware Memory Aggregation
Target Attention

\[ f \left( E_{t_i}^{ta}, E_{r_j}^{ta} \right) = \theta(E_{t_i}^{ta})\varphi(E_{r_j}^{ta})\rho(E_{r_j}^{ta}) \]

Attention between newly extracted embedding and the historical reference of the target.
Distractor Attention

\[ g\left(E_{ti}^{ta}, E_{rj}^{di}\right) = \theta(E_{ti}^{ta})\varphi(E_{rj}^{di})\rho(E_{rj}^{di}) \]

Attention between newly extracted embedding and the historical reference of the closest distractor.
Enhanced Prediction

\[ \tilde{F}_{t}^{ta} = F_{t}^{ta} \oplus w[f(E_{t}^{ta}, E_{r}^{ta}) \ominus g(E_{t}^{ta}, E_{r}^{di})] \]

Enhanced prediction can be performed with \( \tilde{F}_{t}^{ta} \) computed from attentions.
Memory Aggregation

\[
\tilde{E}_{t}^{\text{ta}} = E_{t}^{\text{ta}} \oplus w[f(E_{t}^{\text{ta}}, E_{r}^{\text{ta}}) \ominus g(E_{t}^{\text{ta}}, E_{r}^{\text{di}})]
\]

Raw identity embedding is similarly enhanced by attention modules as \( \tilde{E}_{t}^{\text{ta}} \).

\[
E_{r_{t}} = \text{convGRU}(\tilde{E}_{t}, E_{r_{t-1}})
\]

Then aggregated through convGRU update, and participates in next attention computation

Trained with identity losses
Cross-Task Synergy

(a) Position Prediction by Regression
(b) Temporal-Aware Target Attention & Distractor Attention
(c) Identity-Aware Memory Aggregation
## Benchmark Result

| Method         | MOTA↑ | IDF1↑ | FP↓  | FN↓  | IDS↓ |
|----------------|-------|-------|------|------|------|
| **MOT16**      |       |       |      |      |      |
| MOTD16[10]     | 47.6  | 50.9  | 9253 | 85431| 792  |
| KCF16[12]      | 48.8  | 47.2  | 5875 | 86567| 906  |
| DeepMOT[57]    | 54.8  | 53.4  | 2955 | 78765| 645  |
| Tracktor++V2[3]| 56.2  | 54.9  | **2394** | 76844 | 617  |
| GSM[30]        | 57.0  | 58.2  | 4332 | 73573| **475** |
| TADAM(ours)    | **59.1** | **59.5** | 2540 | **71542** | 529 |
| **MOT17**      |       |       |      |      |      |
| MOTD17[10]     | 50.9  | 52.7  | 24069| 250768| 2474 |
| FANET[13]      | 52.0  | 48.7  | 14138| 253616| 3072 |
| UMA[59]        | 53.1  | 54.4  | 22893| 239534| 2251 |
| DeepMOT[57]    | 53.7  | 53.8  | 11731| 247447| 1947 |
| Tracktor++V2[3]| 56.3  | 55.1  | **8866** | 235449 | 1987 |
| GSM[30]        | 56.4  | 57.8  | 14379| 230174| **1485** |
| TADAM(ours)    | **59.7** | **58.7** | 9676 | **216029** | 1930 |
| **MOT20**      |       |       |      |      |      |
| SORT20[5]      | 42.7  | 45.1  | 27521| 264694| 4470 |
| Tracktor++V2[3]| 52.6  | 52.7  | **6930** | 236680 | 1648 |
| TADAM(ours)    | **56.6** | 51.6  | 39407| **182520** | 2690 |
### Ablation Study Result

| Setup                        | MOTA↑ | IDF1↑ | FP↓  | FN↓  | IDS↓ |
|------------------------------|-------|-------|------|------|------|
| w/o TA & DA                  | 65.9  | 71.1  | 597  | 37501| 208  |
| w/o DA                       | 66.4  | 71.2  | 462  | 37060| 191  |
| w/o TA                       | 66.7  | 71.3  | 473  | 36748| 188  |
| w/o adaptive weight          | 66.0  | 68.5  | 679  | 37322| 242  |
| w/o memory aggregation       | 66.5  | 67.5  | 552  | 36848| 232  |
| Full model                   | 67.0  | 71.6  | 583  | 36287| 197  |
Significantly higher tracked percentage for occlusion level > 50%
Video Result
Thanks for watching!

GitHub Page:
https://github.com/songguocode/TADAM