TrackFlow: Multi-Object Tracking with Normalizing Flows

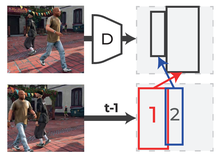
Gianluca Mancusi[1], Aniello Panariello[1], Angelo Porrello[1], Matteo Fabbri[2],
Simone Calderara[1], Rita Cucchiara[1,3]
[1]University of Modena and Reggio Emilia; [2]GoatAI S.r.l.; [3]IIT-CNR, Italy

ICCV23 PARIS

1. Multi-Object Tracking (MOT)

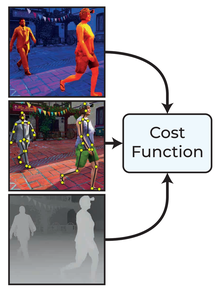
Multi-object tracking. Continuous detection and identification of multiple objects in a video.

Tracking-by-detection. For each frame, candidate bounding boxes are given and associated with tracks.



2. Problem statement

Pose, **depth maps**, or **thermal** data can encode a deeper and robust understanding of the motion and the scene.



How to fuse costs from multiple domains into a single association metric?

Existing works [4, 3] such as Deep SORT exploit handwritten formulas or *heuristics*:

$$c_{i,j} = \lambda\, d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j).$$

The existing fusing strategies:

- assume the costs to be **independent**
- weigh the costs always the same, overlooking *instance*- and *scene*-related peculiarities (*e.g.*, camera motion, lighting conditions, etc.)

Goal. Something like this:

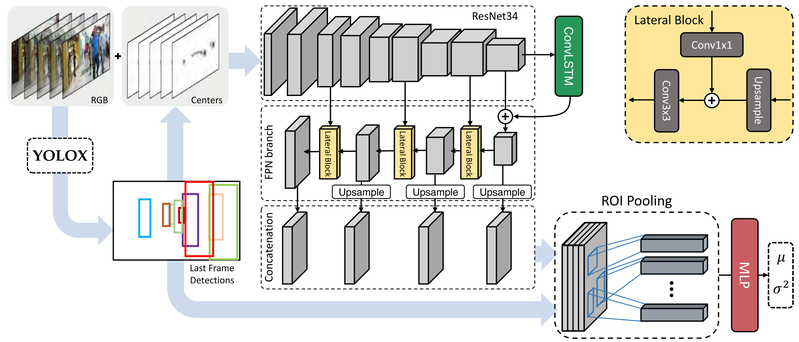
3. DistSynth: estimating per-instance distance from a monocular image

Case study: refining costs related to 2D positions through 3D cues.

DistSynth *(our proposal)*. An ANN giving, for each detected object, its distance from the camera.

It exploits the **Feature Pyramid Network** (FPN) layout to preserve the visual details of distant objects.

It takes the previous $L = 5$ frames as additional input to exploit **temporal patterns** and to handle **temporary occlusions**.



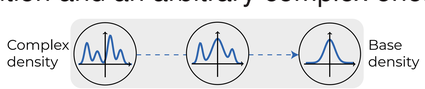
4. Fusing costs: formulation

Formulation. Given the track T, a candidate detection D and the resulting displacements Δ_p, $\Delta_{w,h}$, and Δ_d, we define the fusing cost $\Phi(T, D)$ as the **negative log-likelihood**:

$$\Phi(T, D) = -\log \mathcal{P}_\theta(D \in T \,|\, T).$$

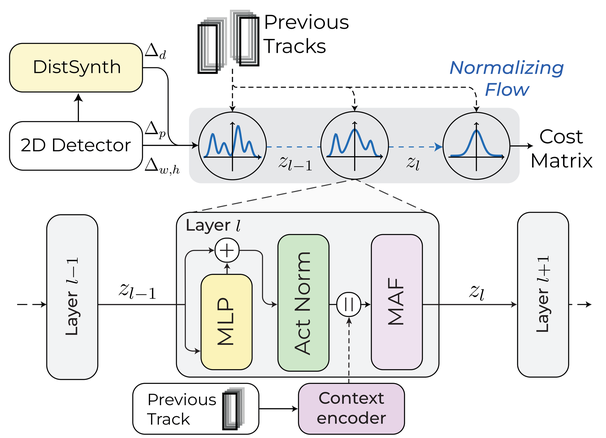
We apply Maximum Likelihood Estimation (MLE) and learn a **deep generative model** $f([\Delta_p, \Delta_{w,h}, \Delta_d] | T, \theta)$ promoting the likelihood of correct associations.

TrackFlow. The design of $f(\cdot | T, \theta)$ derives from **normalizing flow** models, which create an **invertible mapping** between a tractable base distribution and an arbitrary complex one.



- High flexibility
- Exact estimate of the likelihood of a sample

5. TrackFlow



The mapping is carried out through a sequence of L invertible and differentiable transformations:

$$f([\Delta_p, \Delta_{w,h}, \Delta_d] \mid T, \theta) = g_L^{-1} \circ \cdots \circ g_2^{-1} \circ g_1^{-1}.$$

$$\text{forward pass}: \quad \mathbf{z}_l = g_l(\mathbf{z}_{l-1}|T);\ \mathbf{z}_L \sim \mathcal{P}_\theta(D \in T|T)$$

$$\text{inverse pass}: \quad \mathbf{z}_{l-1} = g_l^{-1}(\mathbf{z}_l|T);\ \mathbf{z}_0 \sim \mathcal{N}(0, 1)$$

6. Experiments: multi-object tracking

Results on **MOTSynth** [1] & **MOTChallenge**

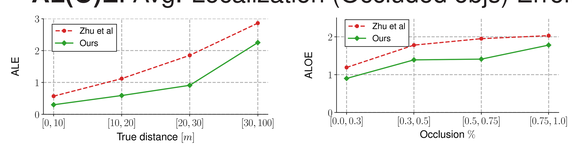
(MOTSynth)	Easy		Moderate		Hard	
Metrics	HOTA ↑	IDF1 ↑	HOTA ↑	IDF1 ↑	HOTA ↑	IDF1 ↑
SORT	63.48	79.40	50.31	62.11	37.48	45.13
+ TrackFlow GT	+4.37	+7.41	+5.33	+9.09	+6.54	+10.88
+ TrackFlow	+0.31	+0.97	+0.81	+1.63	+0.74	+1.56
ByteTrack	63.22	80.84	49.91	62.46	37.61	46.15
+ TrackFlow GT	+3.76	+2.82	+5.47	+5.51	+5.08	+4.60
+ TrackFlow	+0.13	+1.80	+0.47	+1.21	+0.88	+1.81
OC-SORT	65.56	81.61	52.42	63.50	38.10	45.48
+ TrackFlow GT	+2.41	+3.76	+4.88	+7.70	+6.18	+9.55
+ TrackFlow	+0.44	+0.84	+0.60	+1.09	+1.17	+1.96

Improved identity accuracy and steady enhancement of IDF1.

	MOT17		MOT20	
Metrics	HOTA ↑	IDF1 ↑	HOTA ↑	IDF1 ↑
SORT	64.17	72.98	60.56	74.30
+ TrackFlow	+1.78	+1.41	+0.15	+0.22
ByteTrack	67.73	79.81	58.94	74.89
+ TrackFlow	+0.40	+0.23	+0.54	+0.06
OC-SORT	66.22	77.74	55.18	71.22
+ TrackFlow	+0.35	+1.12	+0.53	+0.76

7. Experiments: distance estimation

AL(O)E. Avg. Localization (Occluded objs) Error.



Metrics	$\delta_{<1.25}$ ↑	RMSE ↓	ALP@0.5m	ALP@1m	ALP@2m
SVR	26.7%	12.5	3.4%	6.8%	13.8%
DisNet [2]	27.5%	12.1	3.8%	7.5%	14.6%
Zhu et al. [5]	94.7%	2.15	34.5%	56.2%	78.5%
DistSynth	**99.1%**	**1.91**	**48.0%**	**68.9%**	**86.1%**

[1] M. Fabbri et al. Motsynth: How can synthetic data help pedestrian detection and tracking? In *ICCV*, 2021.
[2] M. A. Haseeb et al. Disnet: a novel method for distance estimation from monocular camera. *10th Planning, Perception and Navigation for Intelligent Vehicles, IROS*, 2018.
[3] J. Rajasegaran et al. Tracking people by predicting 3d appearance, location and pose. In *CVPR*, 2022.
[4] N. Wojke et al. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017.
[5] J. Zhu et al. Learning object-specific distance from a monocular image. In *ICCV*, 2019.