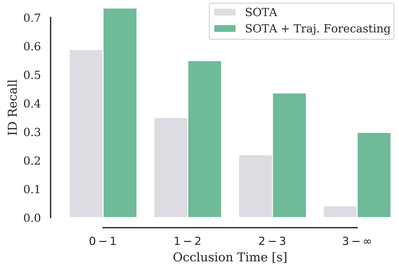
Quo Vadis: Is Trajectory Forecasting the Key Towards Long-Term Multi-Object Tracking?

Patrick Dendorfer *Vladimir Yugay*
Aljoša Ošep *Laura Leal-Taixé*

TL;DR: Quo Vadis leverages pedestrian trajectory prediction in bird's-eye view to solve long-term occlusions in tracking

What problem do we tackle?

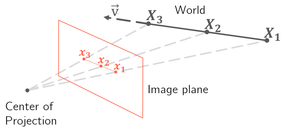
Re-identification after long occlusions is challenging for trackers



Why is reID difficult?

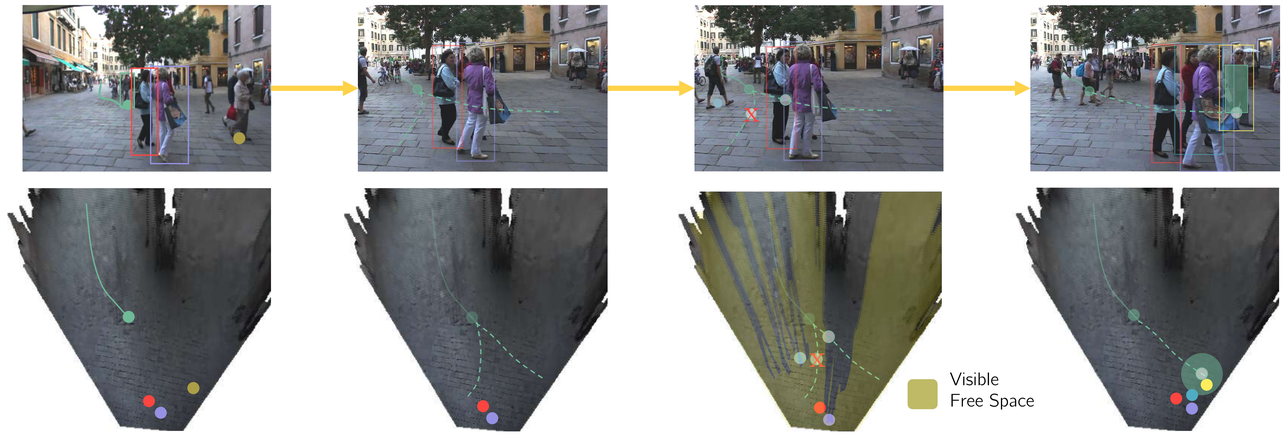
1. Changing Appearance

2. Non-linear camera projection

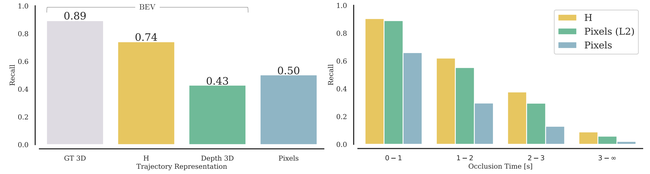


How does the Quo Vadis Pipeline look like?

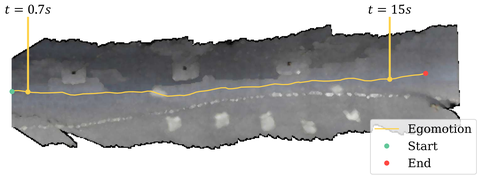
1. Bird's-Eye view of detections and tracks
2. Trajectory forecasting for lost Tracks
3. Filtering incorrect predictions
4. Matching predictions with new detections



Bird's-Eye View Representation



Moving Cameras

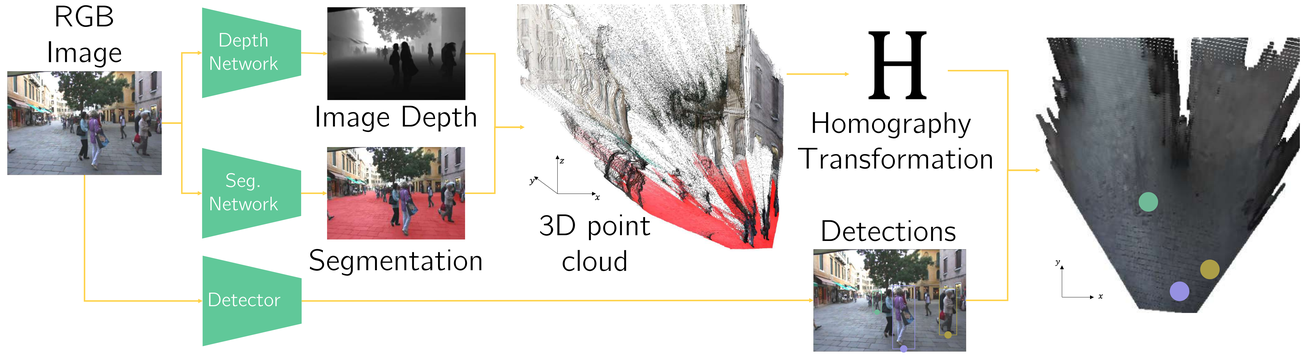


Egomotion estimation for moving cameras

Benchmark Evaluation

	MOT17 (val, static scenes)		MOT17 (val, moving scenes)		MOT20 (train)	
	BYTE	CenterTrack	BYTE	CenterTrack	BYTE	CenterTrack
HOTA	71.36 (+0.21)	61.78 (+3.56)	60.08 (+0.02)	51.77 (+3.07)	56.85 (+0.06)	32.71 (+0.62)
AssA	73.96 (+0.49)	66.18 (+7.54)	60.44 (+0.03)	53.18 (+6.49)	53.97 (+0.20)	28.94 (+1.34)
IDSW	84 (-3)	137 (-146)	54 (+1)	131 (-62)	1815 (-78)	5240 (-2700)
MOTA	80.09 (+0.01)	70.77 (+0.39)	72.54 (-0.01)	59.46 (+0.46)	73.38 (+0.0)	47.57 (+0.24)
IDF1	82.92 (+0.42)	74.46 (+7.13)	73.11 (0.0)	63.48 (+5.76)	72.47 (+0.37)	45.85 (+4.13)

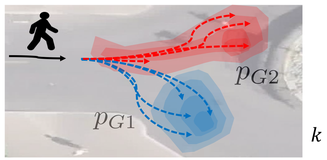
How do we construct the Bird's-Eye View Representation?



Pedestrian Trajectory Prediction

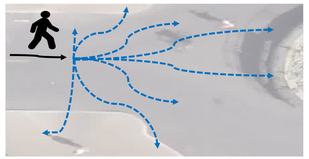
Model	Nr. Samples	Deter-ministic	Stoch-astic	Social	Multi-modal	Prediction $FDE_S \downarrow$	Prediction $FDE_L \downarrow$	Tracking HOTA ↑	AssA ↑	AssRe ↑	AssPr ↑	$ID^{lost}_{IoU} \downarrow$	$ID^{lost}_{\tau} \downarrow$
Baseline	–					–	–	50.71	46.87	51.80	**78.11**	0 %	0 %
Static	1	✓				1.59	2.09	53.84	53.51	60.04	72.95	-14.77 %	-8.40 %
Kalman Filter (pixel)	1	✓				–	–	54.08	54.02	60.45	72.81	**-22.37** %	-8.99 %
Kalman Filter	1	✓				0.69	1.23	54.11	54.04	60.75	71.73	-19.50 %	-16.07 %
GAN	3		✓			0.85	1.26	54.43	54.61	61.11	73.21	-17.99 %	-8.64 %
GAN	20		✓			**0.65**	**0.99**	53.81	53.40	60.45	71.31	-18.03 %	-15.63 %
S-GAN	3		✓	✓		0.87	1.21	**54.52**	54.78	61.22	73.28	-16.92 %	-8.57 %
MG-GAN	3		✓		✓	0.67	1.03	**54.52**	**54.80**	**61.35**	73.13	-21.19 %	**-17.43** %

MG-GAN (multiple generators)



$k = 8$

GAN (single generator)



Track Matching

$$c_{ij} = (\Delta_{\text{IoU}} + \max(\tau_{L_2} - \Delta_{L_2}, 0))$$

with $(\Delta_{\text{App}} \geq \tau_{\text{App}} \text{ and } \Delta_{\text{IoU}} \geq \tau_{\text{IoU}})$

Δ: Distances
τ: Thresholds

IoU: Intersection over Union
App: Appearance Features
L_2: Euclidean Distance

Scores L_2	Scores IoU	Threshold τ_{IoU}	Threshold τ_{App}	HOTA ↑	AssA↑	AssRe ↑	AssPr↑	ID^{lost} ↓
✓				53.89	53.56	60.43	72.21	-16.18 %
✓			✓	53.89	53.57	60.51	71.69	-16.26 %
✓		✓	✓	54.10	53.92	60.43	**73.36**	-16.84 %
	✓			54.13	54.01	60.97	72.00	-24.06 %
✓	✓			53.75	53.35	**61.17**	69.27	**-28.02**%
✓	✓	✓		53.97	53.75	61.08	70.73	-26.93 %
✓	✓		✓	54.06	53.92	61.07	71.01	-21.40 %
✓	✓	✓	✓	**54.27**	**54.29**	61.08	72.36	–20.53%