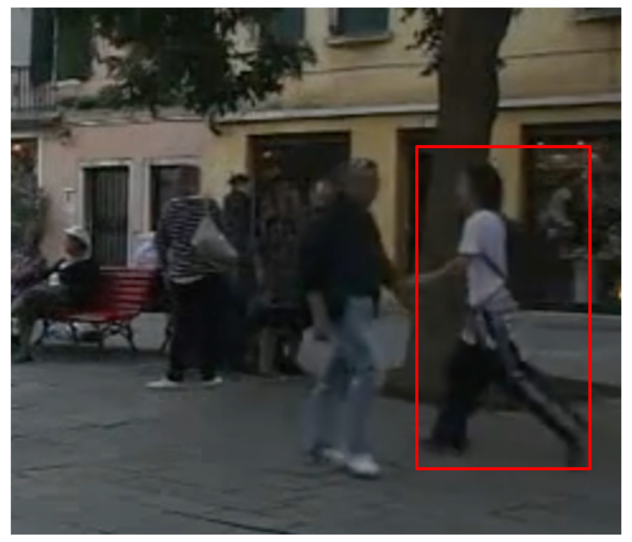
Tracking Objects as Pixel-wise Distributions

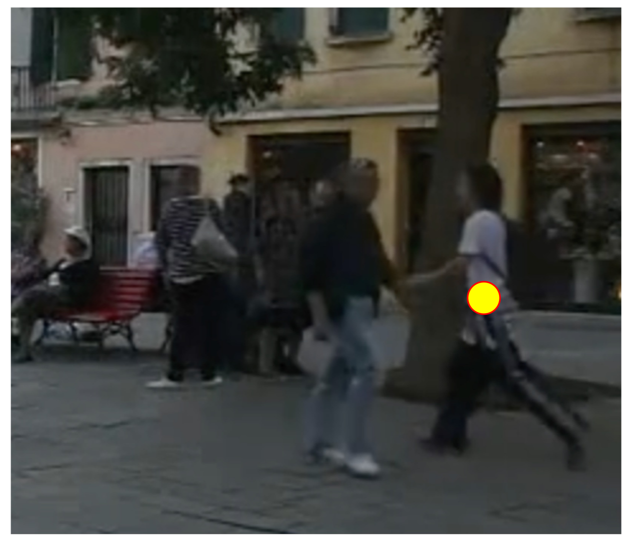
Zelin Zhao [1], Ze Wu [2], Yueqing Zhuang [2], Boxun Li [2], Jiaya Jia [1, 3]
[1] The Chinese University of Hong Kong, [2] MEGVII Technology, and [3] SmartMore

TL; NR: to track objects as pixel-wise distributions.

Bergmann, 2019:
Objects as bounding boxes



Previous 2 (Zhou, 2020):
Objects as points

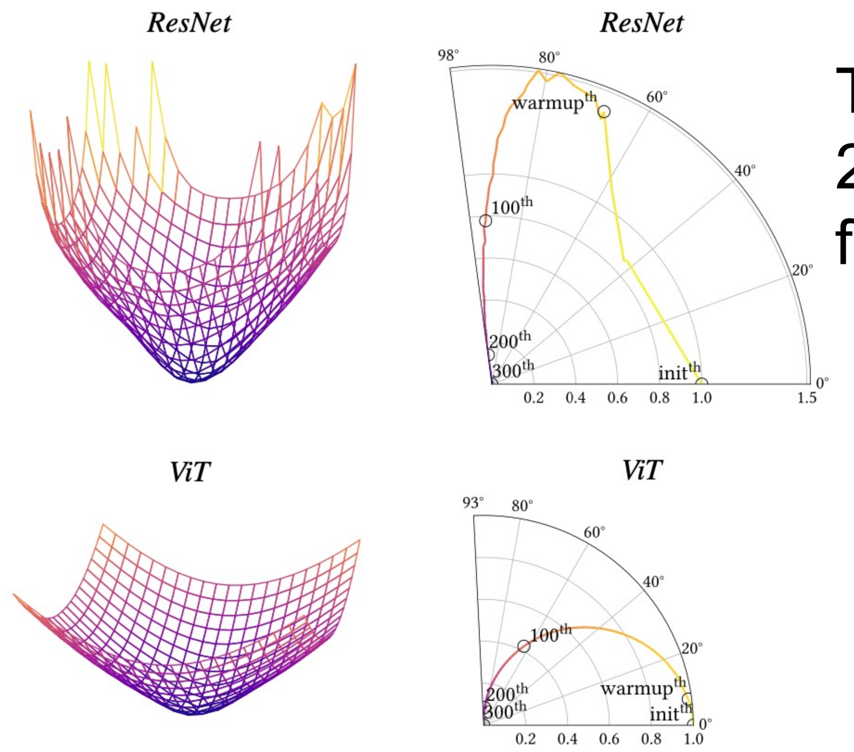


Ours:
objects as pixel-wise distributions

More details, more robust!

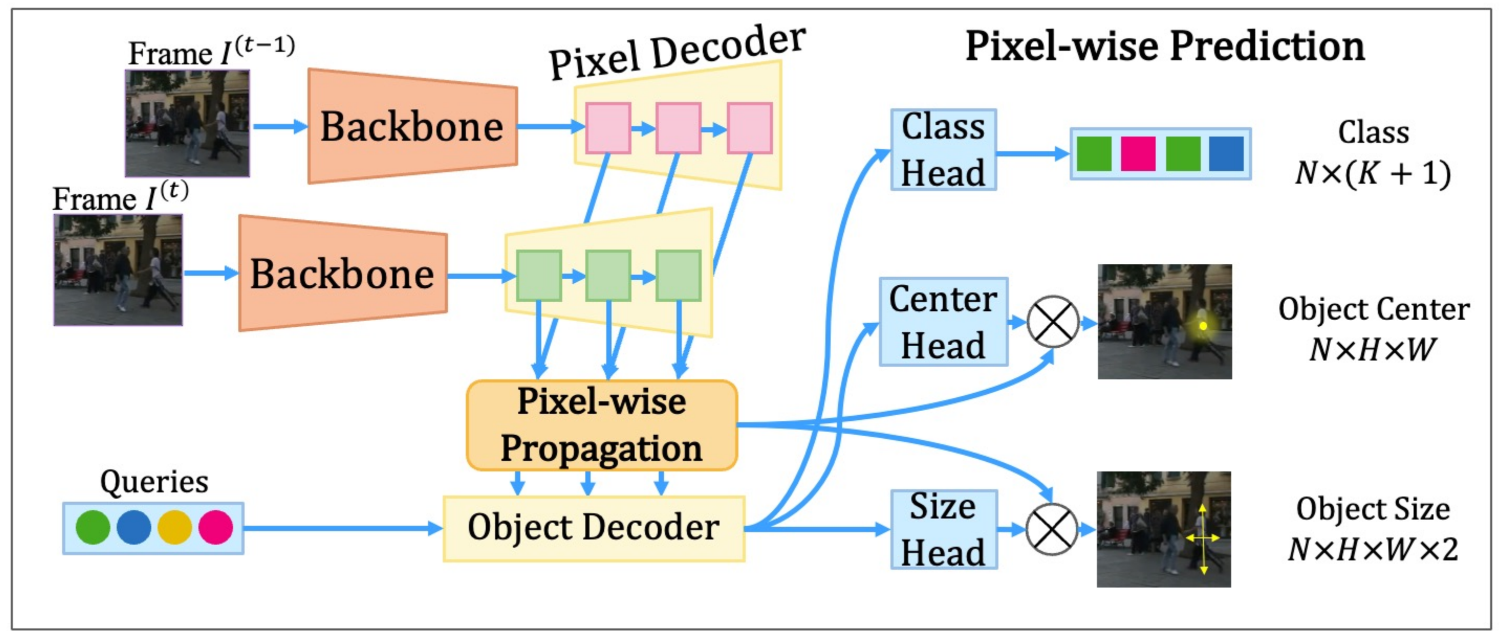
Motivation

1. Pixel-wise information matters (PVNet, Peng, 2018).
2. Low-confident predictions are helpful in MOT (ByteTrack, Zhang, 2021).
3. Smooth prediction leads to better generalization (Park, 2022).



The left image is from (Park, 2022). Transformers benefit from smooth predictions.

P3AFormer model



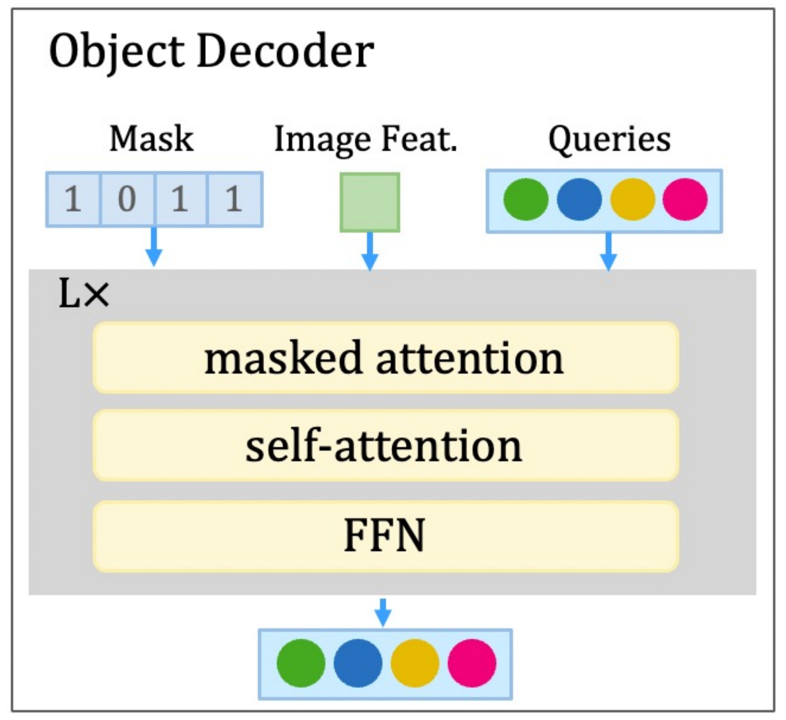
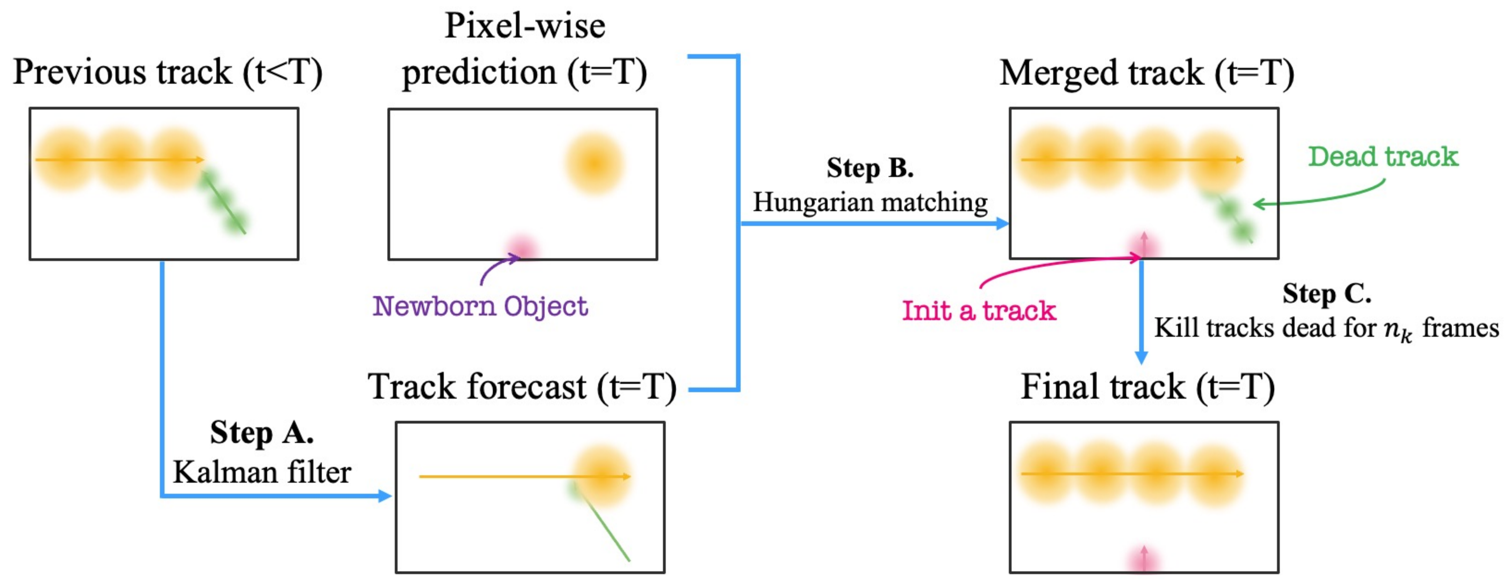


Diagram of P3AFormer model. The backbone encodes the input images, and the pixel decoder produces pixel-level multi-frame feature embeddings. Then the object decoder predicts latent object features, which are passed through several MLP heads to produce class distribution and the pixel-wise representations for object center and size The detailed structure of the object decoder. It uses masked attention, self-attention, and feed-forward networks (FFN) to update the query embedding.



Pixel-wise association scheme in P3AFormer. One object is represented as a pixel-wise distribution, denoted by spheres with the radial gradient change in the above figure. We use one arrow and spheres on the arrow to denote a track.

Results:

Methods	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDSW ↓
FairMOT [76]	73.7	72.3	19.5	36.6	12201	248047	2072
LSST17 [21]	54.7	62.3	20.4	40.1	26091	228434	**1243**
Tracktor v2 [1]	56.5	55.1	21.1	35.3	**8866**	235449	3763
GMOT [29]	50.2	47.0	19.3	32.7	29316	246200	5273
CenterTrack [80]	67.8	64.7	34.6	24.6	18498	160332	3039
QuasiDense [43]	68.7	66.3	40.6	21.9	26589	146643	3378
SiamMOT [49]	65.9	63.3	34.6	23.9	14076	200672	2583
PermaTrack [54]	73.8	68.9	43.8	17.2	28998	115104	3699
CorrTracker [58]	76.5	73.6	47.6	12.7	29808	99510	3369
ByteTrack† [75]	80.3	77.3	53.2	14.5	25491	**83721**	2196
MOTR† [71]	73.4	68.6	42.9	19.1	27939	119589	2439
TransTrack† [51]	74.5	63.9	46.8	**11.3**	28323	112137	3663
TransCenter† [68]	73.2	62.2	40.8	18.5	23112	123738	4614
TransMOT† [11]	76.7	75.1	51.0	16.4	36231	93150	2346
P3AFormer	69.2	69.0	34.8	28.8	18621	152421	2769
P3AFormer (+W&B)	**81.2**	**78.1**	**54.5**	13.2	17281	86861	1893

MOT17 Benchmarks

Visualization of learned pixel-wise distributions.

Methods	mAP ↑	MOTA ↑	IDF1 ↑
w/o All	46.1	71.3	72.1
w/o Mask.	48.0	71.4	74.7
w/o Mix.	47.8	76.6	74.8
w/o Mosiac	46.7	74.0	71.9
w/o LQ	47.9	77.6	75.1
w/o Bbox	48.3	79.1	74.8
with All	48.3	78.4	76.0

Ablation Study of W&B