






BASE: Probably a Better Approach to Visual Multi-Object Tracking

Martin Vonheim Larsen^{1,2,*}^a, Sigmund Rolfsjord^{1,2}^b, Daniel Gusland^{1,2}^c, Jörgen Ahlberg^{3,2}^d
and Kim Mathiassen^{1,2}^e

¹Norwegian Defence Research Establishment, Norway

²University of Oslo, Norway

³Linköping University, Norway
fi

Keywords: Visual Multi-Object Tracking, Probabilistic Tracking, Distance-Aware, Association-Less Track Management.

Abstract: The field of visual object tracking is dominated by methods that combine simple tracking algorithms and ad hoc schemes. Probabilistic tracking algorithms, which are leading in other fields, are surprisingly absent from the leaderboards. We found that accounting for distance in target kinematics, exploiting detector confidence and modelling non-uniform clutter characteristics is critical for a probabilistic tracker to work in visual tracking. Previous probabilistic methods fail to address most or all these aspects, which we believe is why they fall so far behind current state-of-the-art (SOTA) methods (there are no probabilistic trackers in the MOT17 top 100). To rekindle progress among probabilistic approaches, we propose a set of pragmatic models addressing these challenges, and demonstrate how they can be incorporated into a probabilistic framework. We present BASE (Bayesian Approximation Single-hypothesis Estimator), a simple, performant and easily extendible visual tracker, achieving state-of-the-art (SOTA) on MOT17 and MOT20, without using Re-Id. Code available at <https://github.com/ffi-no/paper-base-visapp-2024>.


1 INTRODUCTION


INSTICC:Fx Visual Multi-Object Tracking (VMOT) is the task of estimating the location of objects over time in a video sequence while maintaining a unique ID for each target. Popular VMOT benchmarks (Leal-Taixé et al., 2015; Dendorfer et al., 2020; Sun et al., 2022) are currently dominated by methods which combine simple tracking algorithms with a stack of ad hoc specializations to visual tracking (Zhang et al., 2021; Du et al., 2022; Aharon et al., 2022; Yang et al., 2023). These simple-yet-effective trackers cut corners using hard logic, for instance by ignoring less-certain detections, leaving performance on the table. Meanwhile, *probabilistic trackers* are ubiquitous in more mature fields such as radar- and sonar tracking, as they avoid most of these hard choices and can better


exploit available information. This raises the question: *Why are probabilistic methods outperformed by ad hoc approaches on VMOT?* We believe the probabilistic approaches have overlooked a few key aspects specific to VMOT in their adaptations of existing tracking theory.


In visual tracking, the perspective imaging results in target kinematics and clutter (false detections) characteristics that are radically different from those seen in radar or sonar tracking. When representing target kinematics in image plane coordinates, we should expect objects near the camera to appear more agile than those we see from afar. Similarly, we should expect the density of new targets and clutter detections to be much greater for distant objects, based on the simple fact that objects take up less space in the image when they are farther away. Accounting for these non-uniform effects is necessary to succeed in probabilistic visual tracking.


The dominating ad hoc trackers are built pragmatically from the ground up for VMOT. One example of this is the ubiquitous use of intersection-over-union (IOU) as a similarity metric for detection-to-track matching. Although chosen for its simplicity, the IOU metric has a nice side-effect in that it gives more lee-

^a <https://orcid.org/0000-0002-3008-7712>

^b <https://orcid.org/0009-0004-6118-8593>

^c <https://orcid.org/0009-0001-6351-8128>

^d <https://orcid.org/0000-0002-6763-5487>

^e <https://orcid.org/0000-0003-3747-5934>

*This work was funded by The University Center at Kjeller, and by projects 1505 and 1688 at the Norwegian Defence Research Establishment.

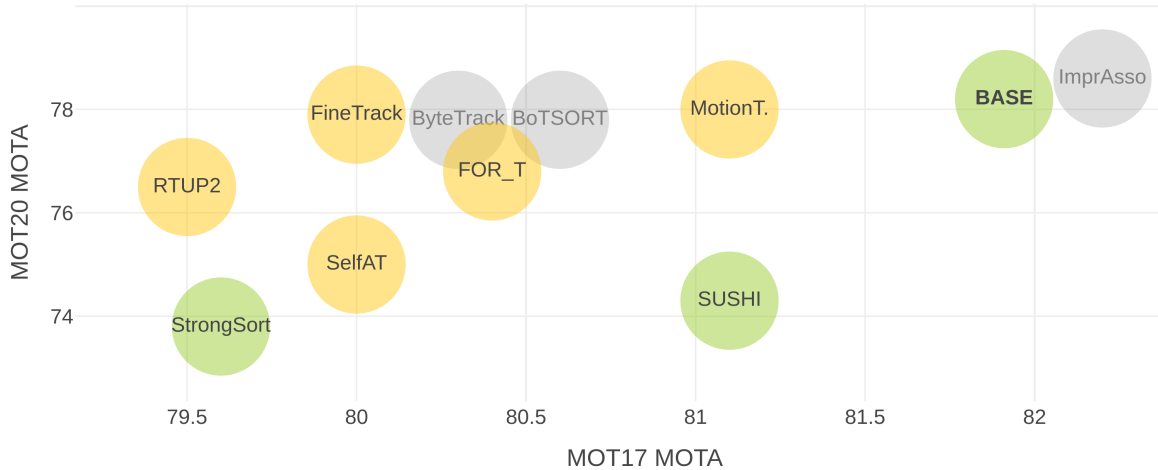


Figure 1: MOTA comparison of our proposed method BASE and top-performing trackers on the MOT17 and MOT20 benchmarks. Trackers in green use a fixed set of parameters across the test set, while those in grey tune parameters for each test sequence. Yellow trackers do not report whether parameters are kept constant on the test set, and have not published code reproducing the results.

way for larger bounding boxes than for smaller ones. Leading approaches (Aharon et al., 2022; Yang et al., 2023) also use object size to change model dynamics when estimating target motion. Avoiding the drawbacks of modeling distance explicitly, these pragmatic mechanisms lead to target models that are *distance-aware*, with good performance both for near and distant objects.

The advantage probabilistic trackers hold over ad hoc methods lies in their ability to better balance different types of information. Which detection is likely to originate from the current object? The one closest in position, the one with the highest confidence, or perhaps the one most similar in size or appearance? Ad hoc methods either ignore parts of this information, or resort to some arbitrary weighting between them. Similarly, ad hoc methods typically accept targets once they have been detected a fixed number of times, ignoring how consistent the detections were, or what confidence level they had. A strong probabilistic approach would instead model the relevant aspects of this information, and take decisions based on what is most *likely*.

Despite the more complex structure of probabilistic trackers, we find that the leading approaches (Fu et al., 2019; Song et al., 2019; Jinlong Yang et al., 2022; Baisa, 2021b) omit, or fail, to model the aspects we see as critical to visual tracking. To lift probabilistic trackers to the performance of the leading ad hoc methods, we propose *Bayesian Approximation Single-hypothesis Estimator (BASE)*, a minimalist probabilistic take on visual tracking. As outlined in Fig. 2, BASE replaces the key components of a traditional single-hypothesis tracker (SHT) with proba-

bilistic counterparts. The novelty of our approach is accounting for the non-uniform kinematics and distribution of clutter experienced in visual tracking, in a unified probabilistic manner. Specifically, our main contributions are:

- An efficient distance-aware motion model.
- Pragmatic modeling of new targets and clutter detections, suited for VMOT.
- A new *association-less* probabilistic track management scheme, suited for crowded scenes.
- Methods for automatically estimating model parameters from training data.

Section 2 gives a brief review of related work, followed by Sec. 3 which revisits the traditional single-hypothesis tracking pipeline. Section 4 describes our proposed method, BASE, and Sec. 5 details our experiments performed on MOT17, MOT20 and DanceTrack. Section 6 contains a summary and conclusions.

2 RELATED WORK

Visual Multi-Object Tracking. The simple and performant trackers that are currently state-of-the-art (SOTA) are mostly based on or inspired by Simple Online and Realtime Tracking (SORT) (Bewley et al., 2016). SORT (Bewley et al., 2016) started a strong line of pragmatic trackers (Wojke et al., 2018; Du et al., 2022; Zhang et al., 2021; Aharon et al., 2022; Cao et al., 2022; Nasseri et al., 2022; Yang et al., 2023; Stadler and Beyerer, 2022). These publications have a strong empirical focus, identifying

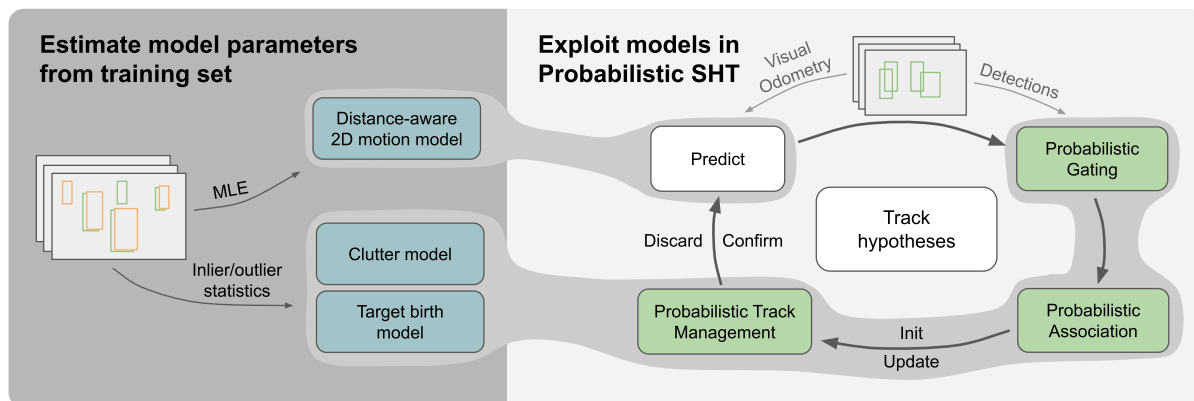


Figure 2: BASE builds on a traditional single-hypothesis tracker (SHT) architecture, but uses probabilistic formulations for all aspects of gating, association and track management, as well as a *distance-aware* motion model. These probabilistic formulations enable BASE to express nuances in detection confidence and detection-to-track match, which threshold-based approaches lack. For a given usecase, BASE requires modeling the motion, clutter, and target birth characteristics of the camera and detection algorithm used, which we can automatically fit using typical training datasets.

more or less standalone components that improve performance. Based on their research, a set of core components have emerged that are common among strong visual trackers. Some of these components are Global Nearest Neighbor (GNN) matching based on IOU-related association metrics, Kalman filtering for track prediction, two-stage track formation, ego-motion compensation, post-interpolation of missed segments, and most importantly, a strong detector.

End-to-End-Trackers: trackers are very attractive, as they can be adapted to new applications without architectural changes. They can also take advantage of information in ways that are hard to express in code. Although many end-to-end trackers exhibit impressive results (Sun et al., 2020; Zhu et al., 2021; Zeng et al., 2021; Yang et al., 2021; Sun et al., 2020; Yan et al., 2022; Wang et al., 2020; Zhang et al., 2023), they lag behind the best tracking-by-detection methods on many VMOT benchmarks like MOTChallenge. This is perhaps due to limited training data and difficulties in training motion models with a short time horizon (Zhu et al., 2021). These issues may be resolved in the future, but we still believe there is room for other approaches, as data restrictions will likely persist in many niche applications.

Probabilistic Visual Tracking. Many probabilistic trackers in the MOTChallenges build on the ideas of Bar-Shalom (Bar-Shalom and Tse, 1975; Bar-Shalom et al., 2007) and Blackman (Blackman and Popoli, 1999). While they mostly rely on advanced state management schemes like MHT (Reid, 1979; Kim et al., 2015) or PHD (Mahler and Martin, 2003), they often resort to non-probabilistic methods for association and for combining different types of information, such as appearance or shape. In fact, the most

successful probabilistic attempts fall back to non-probabilistic association methods like IOU matching. Meanwhile, approaches which rely solely on probabilistic schemes (Baisa, 2021b), score slightly worse than even the extremely simple IOUTracker (Bochinski et al., 2017).

Surprisingly, most probabilistic approaches do not leverage detector confidence beyond basic detection thresholding (Sanchez-Matilla et al., 2016; Fu et al., 2019; Jinlong Yang et al., 2022; Fu et al., 2018; Aguilar et al., 2022; Baisa, 2021b). Some approaches utilize confidence score for track initiation only (Song et al., 2019; Baisa, 2021a; Baisa, 2019). Meanwhile, Wojke et al. (Wojke and Paulus, 2017) demonstrated a significant boost in performance by integrating detector confidence into a PHD filter.

3 PROBABILISTIC SINGLE-HYPOTHESIS TRACKING (SHT) REVISITED

To bridge the gap between the highly specialized ad hoc methods and the overly general probabilistic trackers, the single-hypothesis tracker (SHT) is a good starting point. It is arguably one of the simplest tracking methods that can also be made to leverage most of the key building blocks of probabilistic trackers.

Traditional SHT can be summarized as developing a single set of track hypotheses through the following steps for each new piece of sensor data:

1. **Predict** existing tracks to the current timestep.
2. **Gate** the detections by disregarding detection-to-

track pairs that are considered too unlikely.

3. **Associate** detections to existing track hypotheses, choosing the overall most likely set of matches.
4. **Update** the state estimate of existing tracks using measurement data from the associated detections.
5. **Manage Tracks** by initializing new candidate tracks for unmatched detections, validating promising tracks, and discarding unlikely tracks.

This structure is popular among the top-performing visual trackers, which implement most (or all) these steps using ad hoc modelling (Zhang et al., 2021; Aharon et al., 2022; Du et al., 2022). *Probabilistic SHT* aims to formulate gating, association and track management in terms of probability. These formulations often also require strict probabilistic modeling of the underlying state (predict and update). Blackman (Blackman and Popoli, 1999) gives an excellent in-depth walkthrough of probabilistic SHT, but we will here go through the most critical aspects.

3.1 Probabilistic Gating and Association

The ad hoc trackers discussed in Sec. 2 typically perform gating by thresholding the detection-to-track IOU distance, and association by minimizing the overall IOU distance. The benefit of formulating these mechanisms in terms of probability is that it offers a clear path to incorporate aspects such as the quality of the track estimates, the detector performance, and the clutter characteristics.

For a given true track state \mathbf{x}_i , we assume to have an estimate $\hat{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{x}_i, \mathbf{P}_i)$ and that corresponding detections \mathbf{z}_j are generated as

$$\mathbf{z}_j = \mathbf{H}\mathbf{x}_i + \mathbf{w}_j, \mathbf{w}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_j). \quad (1)$$

Here, \mathbf{H} is the measurement function, \mathbf{w}_j is white measurement noise, and \mathbf{P}_i and \mathbf{R}_j are known covariances. We can then consider the detection-to-track innovation

$$\hat{\mathbf{y}}_{ij} = \mathbf{z}_j - \mathbf{H}\hat{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{S}_{ij}), \quad (2)$$

with $\mathbf{S}_{ij} = \mathbf{R}_j + \mathbf{H}\mathbf{P}_i\mathbf{H}^\top$.

For gating, an traditional approach (see (Blackman and Popoli, 1999, Sec. 6.3.2) for details) is to threshold the match-to-noise likelihood ratio:

$$\frac{\mathcal{N}(\hat{\mathbf{y}}_{ij}; \mathbf{0}, \mathbf{S}_{ij})}{\lambda_C + \lambda_{NT}} \geq \frac{1 - P_G}{P_G}, \quad (3)$$

where λ_{NT} and λ_C are the densities of new targets and clutter, and P_G is the desired gating confidence level.

For probabilistic association, we want to find the “most likely” set of detection-to-track pairs $\mathcal{A} =$

$\{(j, i)\}$. A common approach, which ignores the detector performance and clutter characteristics, is to formulate this as linear-sum assignment problem over the negative logarithm of the innovation likelihood:

$$\mathcal{A}^* = \arg \min_{\mathcal{A}} \sum_{(j,i) \in \mathcal{A}} \hat{\mathbf{y}}_{ij}^\top \mathbf{S}_{ij}^{-1} \hat{\mathbf{y}}_{ij} + \log |\mathbf{S}_{ij}| \quad (4)$$

3.2 Probabilistic Track Management

Where ad-hoc trackers employ schemes such as counting the number of recent detections to assess whether a track hypothesis should be discarded, probabilistic SHT estimates the probability that each hypothesis is valid. This is done by assessing the event

\mathcal{H}_i : x_i was either not detected, or explained by a detection from a real target, in each frame.

For each track i we then maintain a likelihood-ratio (LR) LR_i , weighing evidence for- and against \mathcal{H}_i , as

$$\text{LR}_i \triangleq \frac{p(\mathbf{Z}_k | \mathcal{H}_i)}{p(\mathbf{Z}_k | \overline{\mathcal{H}}_i)} \dots \frac{p(\mathbf{Z}_1 | \mathcal{H}_i)}{p(\mathbf{Z}_1 | \overline{\mathcal{H}}_i)} \frac{\Pr\{\mathcal{H}_i\}}{\Pr\{\overline{\mathcal{H}}_i\}}, \quad (5)$$

where \mathbf{Z}_k represents the sensor data at time k , and $\overline{\mathcal{H}}_i$ is the logical complement of \mathcal{H}_i . At each timestep k , depending on whether the track was detected or not, Blackman writes the corresponding LR factor as

$$\begin{aligned} & \frac{p(\mathbf{Z}_k | \mathcal{H}_i)}{p(\mathbf{Z}_k | \overline{\mathcal{H}}_i)} \\ & \triangleq \begin{cases} P_D \frac{p(\hat{\mathbf{y}}_{ij})}{\lambda_C} \frac{p(z_S | \text{Det}, \mathcal{H}_i)}{p(z_S | \text{Det}, \overline{\mathcal{H}}_i)}, & \text{if assoc. to } j \\ 1 - P_D, & \text{otherwise.} \end{cases} \end{aligned} \quad (6)$$

Here, P_D is prior detection probability, which is typically modelled as a constant. $\text{LR}_S = \frac{p(z_S | \text{Det}, \mathcal{H}_i)}{p(z_S | \text{Det}, \overline{\mathcal{H}}_i)}$ is

the “signal-related” LR, typically derived from the SNR of the given detection method. We then consider a track hypothesis x_i to be “unconfirmed” until LR_i passes some given threshold, and discard the hypothesis if LR_i falls below some other threshold.

4 THE BASE VISUAL TRACKER

In this section, we present BASE, a minimalist probabilistic take on visual tracking. We design BASE as a probabilistic SHT with the necessary extensions to sufficiently model the visual tracking problem, as shown in Figure 2.

To accommodate the non-uniform motion and clutter encountered in visual tracking, we develop

the *distance-aware 2D motion model* in Sec. 4.4 and model detector performance and clutter behaviour in Sec. 4.5. In Sec. 4.6 we propose an automatic procedure to estimate the parameters of these models. The traditional SHT pipeline discussed in Sec. 3 cannot fully exploit these models in all aspects of gating, association and track management. We therefore compute the *association probability* in Sec. 4.1, before using it to reformulate gating and association in Sec. 4.2 and track management in Sec. 4.3.

4.1 Explicitly Approximating the Association Probability

To account for non-uniform densities of new targets and clutter, the traditional association score from Eq. (4) is insufficient. Consider a track which is presented with two measurements that have identical statistical distance, but where one measurement has far greater risk of being clutter than the other. Intuitively, we should prefer to associate the track with the measurement least likely to be clutter.

Instead of using the traditional association score from Eq. (4), we shall compute the full *association probability* for each detection/track pair. First, we define the event

\mathcal{A}_{ij} : z_j originates from the real target represented by x_i .

In our single-hypothesis paradigm, each detection z_j must originate either from a target for which we have an hypothesis, a newly appeared target, or from clutter. Given the corresponding measurement \mathbf{z}_j we use

$$\lambda_{\text{EX}}(\mathbf{z}_j) = \lambda_{\text{NT}}(\mathbf{z}_j) + \lambda_{\text{C}}(\mathbf{z}_j) \quad (7)$$

to denote the corresponding *extraneous measurement density*. Modelling λ_{NT} and λ_{C} is detailed in Sec. 4.5.

We can now write the association probability as

$$\Pr\{\mathcal{A}_{ij} | \hat{\mathbf{x}}_i, \mathbf{z}_j\} = \frac{p(\hat{\mathbf{x}}_i, \mathbf{z}_j | \mathcal{A}_{ij})}{\lambda_{\text{EX}}(\mathbf{z}_j) + \sum_l p(\hat{\mathbf{x}}_l, \mathbf{z}_j | \mathcal{A}_{lj})}. \quad (8)$$

Here, $p(\hat{\mathbf{x}}_i, \mathbf{z}_j | \mathcal{A}_{ij})$ is the joint likelihood of the current track state $\hat{\mathbf{x}}_i$ and the observed measurement \mathbf{z}_j , assuming that z_j originates from x_i . Inspired by (Blackman and Popoli, 1999) we model our measurements to consist of a *state-related measurement* (modelled in Sec. 4.4) and an independent *confidence measurement* (modelled in Sec. 4.5). Since we will only model the bounding box state of each track, we write

$$p(\hat{\mathbf{x}}_i, \mathbf{z}_j | \mathcal{A}_{ij}) \triangleq p(\hat{\mathbf{x}}_i^{(\text{bb})}, \mathbf{z}_j^{(\text{bb})} | \mathcal{A}_{ij}) p(\mathbf{z}_j^{(\text{c})}). \quad (9)$$

4.2 Probabilistic Gating and Association

Instead of using the traditional Eq. (3) for gating and Eq. (4) for association scores, we will base both gating and scoring on Eq. (8). We compute association scores as

$$-\log \Pr\{\mathcal{A}_{ij} | \hat{\mathbf{x}}_i, \mathbf{z}_j\} \quad (10)$$

and perform gating using

$$\Pr\{\mathcal{A}_{ij} | \hat{\mathbf{x}}_i, \mathbf{z}_j\} \geq \frac{1 - P_G}{P_G}. \quad (11)$$

The introduction of Eq. (8) results in stricter than before gates both for measurements in crowded regions and less confident measurements. For association, however, Eq. (8) also takes the extraneous measurement density into account, critical to properly balance between measurements that have vastly different λ_{EX} . Using Eq. (10), the linear-sum assignment will now find the set of associations with the overall lowest probability of containing a misassociation, whereas the traditional variant (Eq. (4)) finds the most likely association only with regards to the predicted state vs the observed measurements.

4.3 Probabilistic Track Management

In BASE, we will build the track management around the probability that *at least one of the measurements originates from a given track x_i* in the current frame, namely

$$\tilde{P}_i = 1 - \prod_j (1 - \Pr\{\mathcal{A}_{ij} | \hat{\mathbf{x}}_i, \mathbf{z}_j\}). \quad (12)$$

To rewrite Eq. (5) using \tilde{P}_i , we first define the event

$$\mathcal{D}_i^{(k)} : x_i \text{ was detected in frame } k. \quad (13)$$

Re-ordering Eq. (5) with Bayes' rule, we can write the LR-contribution of frame k as

$$\begin{aligned} & \frac{p(\mathcal{H}_i | Z_k)}{p(\overline{\mathcal{H}}_i | Z_k)} \\ &= \frac{p(\mathcal{H}_i, \mathcal{D}_i^{(k)} | Z_k) + p(\mathcal{H}_i, \overline{\mathcal{D}}_i^{(k)} | Z_k)}{p(\overline{\mathcal{H}}_i, \mathcal{D}_i^{(k)} | Z_k) + p(\overline{\mathcal{H}}_i, \overline{\mathcal{D}}_i^{(k)} | Z_k)} \end{aligned} \quad (14)$$

$$= \frac{\tilde{P}_i^{(k)} + (1 - P_D)(1 - \tilde{P}_i^{(k)})}{P_D(1 - \tilde{P}_i^{(k)})}, \quad (15)$$

where we have used that $\mathcal{D}_i^{(k)} \subset \mathcal{H}_i$, and that

$$P_D \triangleq \Pr \left\{ \mathcal{D}_i^{(k)} \mid \mathcal{H}_i \right\}. \quad (16)$$

Where the traditional LR update (Eq. (6)) only uses the associated measurement, Eq. (15) collects contributions from *all* measurements, normalized across all tracks. This *association-less* track LR better handles the cases where several tracks have significant stakes in a given measurement, which is often the case in crowded visual tracking, as illustrated in Fig. 3. Since this computation is well suited for GPU acceleration and can be run in parallel with the association, the increased computational burden is more than made up for in practice.

4.4 A Distance-Aware Planar Motion Model for Bounding-Box Objects

We shall stick to a traditional linear Kalman filter (KF) setup to model measurements and target motion. For each target, we model the bounding box center, width and height in the current image plane using pixel coordinates. We assume a nearly constant velocity (NCV) model for the bounding box center c_x, c_y , and use a nearly constant state model for the box size. As a state vector we choose

$$\mathbf{x} = (c_x, c_y, \dot{c}_x, \dot{c}_y, w, h), \quad (17)$$

where \dot{c}_x, \dot{c}_y denote the center velocity, and w and h is the width and height. We model each NCV block using

$$\mathbf{F}_{cv} = \begin{bmatrix} 1 & \delta_t \\ 0 & 1 \end{bmatrix}, \mathbf{Q}_{cv} = \begin{bmatrix} \delta_t^3/3 & \delta_t^2/2 \\ \delta_t^2/2 & \delta_t \end{bmatrix}, \quad (18)$$

for a given timestep δ_t , as in (Blackman and Popoli, 1999, Sec. 4.2.2). We use \mathbf{F}_{cv}^{*2} and \mathbf{Q}_{cv}^{*2} to denote the 2D composition of \mathbf{F}_{cv} and \mathbf{Q}_{cv} .

We correct for camera ego-motion by first aligning each new image to the previous one using (Evan gelidis and Psarakis, 2008). Between each pair of following images we obtain a transform which predicts pixels in time as

$$\mathbf{p}_k = \mathbf{W}_k \mathbf{p}_{k-1} + \mathbf{t}_k. \quad (19)$$

By using the notation $\mathbf{T}_k = \text{diag}(\mathbf{W}_k, \mathbf{W}_k, \mathbf{W}_k)$ and $\mathbf{F}_k = \text{diag}(\mathbf{F}_{cv}^{*2}, \mathbf{I}_2)$, we can write the state transition as

$$\mathbf{x}_k = \mathbf{T}_k (\mathbf{F}_k \mathbf{x}_{k-1} + \mathbf{v}_k) + \mathbf{I}_{6 \times 2} \mathbf{t}_k, \quad (20)$$

with white $\mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$, where

$$\mathbf{Q}_k = \sigma_k^\top \text{diag}(\mathbf{Q}_{cv}^{*2}, \mathbf{I}_2) \sigma_k. \quad (21)$$

The key to making the model distance-aware, is here

to scale σ_k with the previous object width, w_{k-1} , similar to what is done in (Aharon et al., 2022):

$$\sigma_k = w_{k-1} (\sigma_{ca}, \sigma_{ca}, \sigma_{ca}, \sigma_{ca}, \sigma_{sr}, \sigma_{sr}), \quad (22)$$

where σ_{ca} and σ_{sr} are the standard deviation of the center acceleration noise and the size rate noise, respectively.

For the sensor model we assume that we make observations corrupted by white Gaussian noise as

$$\mathbf{z}_k = (\mathbf{x}_k)_{c_x, c_y, wh} + \mathbf{w}_k, \mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k). \quad (23)$$

The sensor and transition models together describe a linear Gaussian system, which is suitable for estimation using a KF. When initializing new tracks, we will in addition to the above employ

$$\hat{\mathbf{x}}_0 = (\mathbf{z}, 0, 0) \quad (24)$$

$$\hat{\mathbf{P}}_0 = \text{diag}(\mathbf{R}, \mathbf{P}_{cr}) \quad (25)$$

as a prior for the center rate, where \mathbf{R} and \mathbf{P}_{cr} are the to-be-estimated measurement- and initial center rate covariances.

4.5 Modelling Detector Performance

In probabilistic tracking, the extraneous measurement density (λ_{EX} from Eq. (7)) and the detection confidence ($p(\mathbf{z}_j^{(c)})$ from Eq. (9)) are typically ignored or treated as constants. However, these quantities describe critical aspects of the detector performance that should affect the tracking. Properly modeling these parameters can allow us to quickly establish track in the simple cases, while still avoiding false tracks in questionable cases.

We begin by modeling the detector confidence $\mathbf{z}^{(c)}$ in Eq. (9) through histogram binning of inlier and all detections on the training set (see Sec. 4.6 for details) as

$$p(\mathbf{z}_j^{(c)}) \triangleq \frac{\text{hist}_{\text{inlier}}(c_j, w_j)}{\text{hist}_{\text{all}}(c_j, w_j)}, \quad (26)$$

where w_j is the measured width, and c_j is the predicted confidence from the detector. Figure 5 shows the resulting likelihood on MOT17, which is clearly nonuniform.

Next, we shall model $\lambda_{EX}(\mathbf{z})$, which is the density of clutter measurements and measurements due to newly appeared targets. Through experimentation, we have found that object size alone is a good discriminator for λ_{EX} . We therefore use scaled histogram binning over object width across *all* training set detections as a pragmatic model:

$$\lambda_{EX}(\mathbf{z}_j) \triangleq c_{EX} \text{hist}_w(w_j), \quad (27)$$



Figure 3: Our proposed association-less track management from Eq. (15) vs. traditional association-based track management on “MOT20-07”. Select tracks are highlighted in color based on track-ID, while boxes of unrelated tracks are drawn in gray. In crowded areas, the detector tends to generate clutter detections, which can lead to false tracks, as in Fig. 3b. Meanwhile, detections of partially occluded objects tend to have low confidence, which can cause the tracker to require more frames before a track is established, as in Fig. 3d. Our association-less track management enables the tracker to be conservative in establishing tracks in crowded areas, and at the same time aggressive for solitary objects. Meanwhile, association-based track management must more carefully balance the risk of false tracks and delayed track establishment.

where w_j is the measured width and c_{EX} is a constant we estimate in Sec. 4.6. Figure 6 shows a log-plot of hist_w for MOT17, which is clearly skewed towards smaller boxes.

4.6 Automatically Estimating Model Parameters from Training Data

To start using the proposed motion and sensor model, we need the parameters σ_{ca} , σ_{sr} , \mathbf{R} and \mathbf{P}_{cr} . Fortunately, these can be estimated from a dataset consisting of detections and ground-truth tracks, such as those provided in the MOTChallenges.

The ground truth bounding boxes are given as $\mathbf{g}_i^{(k)} = (c_x, c_y, w, h)$ for each true target x_i in each frame k where x_i is present. We start with the prior center rate covariance \mathbf{P}_{cr} , which can be estimated from the ground truth tracks alone. To avoid potential errors in the camera ego-motion correction tainting the ground truth data, we only use sequences where the camera is stationary. We then estimate \mathbf{P}_{cr} as

$$\mathbf{P}_{cr} = \frac{1}{n_g - 1} \sum_i \frac{(\mathbf{g}_i^{(k_2)} - \mathbf{g}_i^{(k_1)})(\mathbf{g}_i^{(k_2)} - \mathbf{g}_i^{(k_1)})^\top}{(t_i^{(k_2)} - t_i^{(k_1)})^2}, \quad (28)$$

where k_1 and k_2 index the frames where target i ap-

pears for the first and second time, $t_i^{(k_2)} - t_i^{(k_1)}$ is the time elapsed between said frames, and n_g is the total number of targets.

The detection bounding boxes are given as a set $\mathbf{z}_j = (\hat{c}_x, \hat{c}_y, \hat{w}, \hat{h}) \in \mathcal{Z}$ for each frame. To leverage the detections in parameter estimation, we first attempt to associate detections \mathbf{z}_j to ground truth targets \mathbf{g}_i using IOU. We only consider pairs where $\text{iou}(\mathbf{z}_j, \mathbf{g}_i) > 0.7$, and match \mathbf{z}_j to \mathbf{g}_i if \mathbf{z}_j is the closest to \mathbf{g}_i and vice-versa. We denote the resulting set of associations $\mathcal{A} = \{(\mathbf{z}_{j_i}, \mathbf{g}_{i_i})\}_I$.

Using the associated detections and ground truth targets, we once again use stationary sequences, and estimate \mathbf{R} as

$$\mathbf{R} = \frac{1}{n_a - 1} \sum_{(\mathbf{z}_j^{(k)}, \mathbf{g}_i^{(k)}) \in \mathcal{A}_k} (\mathbf{z}_j^{(k)} - \mathbf{g}_i^{(k)})(\mathbf{z}_j^{(k)} - \mathbf{g}_i^{(k)})^\top, \quad (29)$$

where n_a is the total number of associations.

The histograms over all detections as a function of box width ($\text{hist}_w(w)$) and as a function of both predicted confidence and box width ($\text{hist}_{\text{all}}(c, w)$) can be computed directly from the training set. We compute inlier histogram ($\text{hist}_{\text{all}}(c, w)$) using only the associated detections in \mathcal{A} .

To estimate σ_{ca} and σ_{sr} we employ maximum likelihood estimation (MLE) based on \mathcal{A} and the pro-

posed motion- and sensor models from Sec. 4.4, as outlined in (Brekke, 2019). Finally, we find c_{EX} by a parameter search where we run the tracker on the full training set.

5 EXPERIMENTS

To assess the effectiveness of our proposed minimalist probabilistic visual tracker, we evaluate BASE on the MOT17 (Milan et al., 2016), MOT20 (Dendorfer et al., 2020) and DanceTrack (Sun et al., 2022) benchmarks. Since we are primarily interested in validating the probabilistic backbone, we opt not to use Re-Id or other appearance features. We will focus on the MOTChallenge benchmarks, which are ideal for demonstrating a minimal probabilistic visual tracker as they contain single-camera footage with simple ego-motion and no complex movement patterns. The DanceTrack dataset contains much more sudden movement of arms and legs, and would benefit from a more specialized motion model.

5.1 MOTChallenge Caveats

Although the MOTChallenge benchmarks enable objective comparison of tracking algorithms, there are a few noteworthy differences in practices that color the results.

Not All Results Use Global Parameters. As discussed in (Cetintas et al., 2023), several submissions boost performance by tuning separate parameter sets for each sequence in the *test set*. Unless explicitly stated or evident from published source code, we cannot ascertain which practice is used for a given method.

Post-Tracking Interpolation across frames where objects are not observed, is ubiquitous among all top-scoring methods on both MOT17 and MOT20. Approaches that interpolate over a fixed number of frames, still seem to consider themselves “online”.

The public Detection Leaderboard seems useful to compare trackers on equal terms. However, all the top submissions in this category still use image data to extract additional detections or Re-ID, greatly occluding the results.

5.2 Detector

We use the YOLOX (Ge et al., 2021) detector with a detection threshold of 0.1 for all benchmarks. For DanceTrack we use the officially trained weights. For MOT17 and MOT20 we trained the model in a similar fashion as ByteTrack (Zhang et al., 2021), with



Figure 4: An image from “MOT17-05-FRCNN” processed using the pretrained detector from ByteTrack (Zhang et al., 2021). This detector is overfitted to “see through” severe occlusions, resulting in a large number of clutter detections.

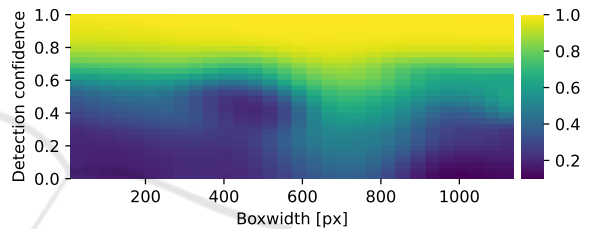


Figure 5: The $p(\mathbf{z}_j^{(c)})$ histogram from Eq. (26) for the MOT17 training set. Each cell is the number of inlier detections (as defined in Sec. 4.6) divided by the number of detections within the corresponding boxwidth/score bin.

a combination of MOT17, Cityperson (Zhang et al., 2017), Crowdhuman (Shao et al., 2018) and ETHZ (Ess et al., 2008). However, as illustrated in Fig. 4, the original ByteTrack detector has been severely overfitted to propose detections even for occluded objects. We therefore excluded fully occluded and *crowd* targets in the training process to avoid this type of overfitting.

5.3 Parameter Estimation

For each of the three benchmarks we select a global set of parameters that are used across all sequences. Following the scheme outlined in Secs. 4.5 and 4.6 we compute \mathbf{P}_{cr} , \mathbf{R} , $p(\mathbf{z}^{(c)})$ and histograms. σ_{ca} and σ_{sr} are estimated using MLE on inlier detections, and c_{EX} is found by a parameter search using full tracking on the respective training sets.

For all benchmarks we use $P_D = 0.95$. MOT17 and MOT20 use a canonical $P_G = 10^{-3}$. DanceTrack has very few distracting elements, and we have seen improved performance using $P_G = 10^{-6}$.

Table 1: SOTA and select methods on key benchmarks, sorted by MOT17 MOTA. The column with grey background shows top results using strictly public detections on MOT17. Results which tune parameters for each sequence in the test set are shown in grey font.

Method	ReID	MOT17		MOT20		Dancetrack		MOT17 Public	
		MOTA	HOTA	MOTA	HOTA	MOTA	HOTA	MOTA	HOTA
PHD_GM (Sanchez-Matilla et al., 2020)	-	-	-	-	-	-	-	48.8	-
GMPHDO. (Song et al., 2019)	-	-	-	-	-	-	-	49.9	-
OCSORT (Cao et al., 2022)	-	78.0	63.2	75.7	62.4	92.0	55.7	-	-
MOTRv2 (Zhang et al., 2023)	✓	78.6	62.0	76.2	60.3	92.1	73.4	-	-
StrongSort (Du et al., 2022)	✓	79.6	64.4	73.8	62.6	-	-	-	-
FineTrack (Ren et al., 2023)	✓	80.0	64.3	77.9	63.6	89.9	52.7	-	-
ByteTrack (Zhang et al., 2021)	-	80.3	63.1	77.8	61.3	90.9	51.9	-	-
BoT-SORT (Aharon et al., 2022)	✓	80.5	65.0	77.8	63.3	-	-	-	-
SUSHI (Cetintas et al., 2023)	✓	81.1	66.5	74.3	64.3	88.7	63.3	-	-
MotionTrack (Qin et al., 2023)	✓	81.1	65.1	78.0	62.8	-	-	-	-
CBIoU (Yang et al., 2023)	✓	81.1	64.1	-	-	91.6	60.6	-	-
ImprAsso (Stadler and Beyerer, 2023)	✓	82.2	66.4	78.6	64.6	-	-	-	-
BASE (ours)	-	81.9	64.5	78.2	63.5	91.7	56.4	51.8	43.6

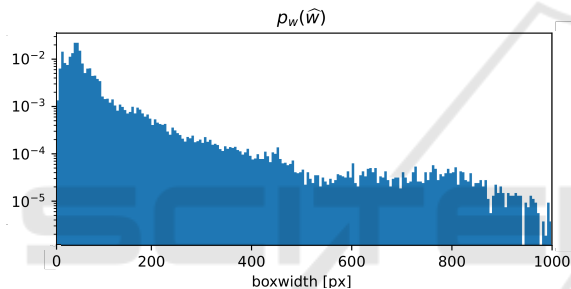


Figure 6: The $\text{hist}_w(w_j)$ histogram from Eq. (27) for the MOT17 training set. Each cell is just the density of detections within the corresponding boxwidth bin.

5.4 Post-Processing

The MOTChallenge ground truth used for scoring contains both visible and occluded targets, which makes it important to maintain tracks through occlusions. True real-time trackers will always have a disadvantage in this regard, as they cannot correct already reported trajectories when old targets reappear. To counter this, it has become a common practice among the top performing tracker to apply interpolation across such occlusion gaps in a post-processing step. With this post-processing the tracker can still run live, but will have some fixed delay.

We employ the interpolation post-processing as described in (Zhang et al., 2021), in addition to what we call *look-ahead*. With look-ahead we delay reporting of tracks by a fixed number of frames, but use the estimated track hypothesis likelihood of the newest processed frame to determine whether the track should be reported. The primary benefit of this is that we can report tracks with certainty upon first detection, even though the track hypothesis likelihood

requires a few frames to accumulate. Without look-ahead, the metrics used in the MOTChallenges force the tracker to establish tracks on the very first detections, severely limiting the ability of the probabilistic track management to filter clutter.

5.5 MOT17 Results

Our overall results for MOT17 are presented in Fig. 1 and Tab. 1. On the testset, BASE (excl. the detector) ran at 331Hz on an AMD 5950x. Among all submissions with publications and which use global parameters, BASE ranks first with an 81.9 MOTA score and third with a 64.5 HOTA score. In MOTA, BASE is only surpassed by ImprAsso (Stadler and Beyerer, 2023), which transparently report using per-sequence tuning on the test set. We strongly suspect that BASE would receive a significant performance boost using a similar tuning scheme, but we insist on using global parameters. Compared to leading trackers, like BoT-SORT (Aharon et al., 2022), BASE performs exceedingly well on sequences with small objects and persistent clutter detections, like sequence 01 (72.4 vs 63.4 MOTA) and 14 (68.1 vs 53.5 MOTA). We believe this owes to BASE’s probabilistic track management, which better captures the nuance between faint detections of small objects and more inconsistent clutter detections.

Compared to BoT-SORT, BASE performs worse on sequences with long occlusions paired with large camera motion, like sequence 06 (64.8 vs 66.6 MOTA). We suspect this is due to our lack of Re-ID, as other motion-only approaches, such as ByteTrack and OC-SORT, also perform poorly on this sequence (60.2 and 57.3 MOTA).

Table 2: Ablation results on the MOT17 validation set. The experiments quantify the effects of our proposed dynamic clutter density model, the distance-aware motion model, the probabilistic association and the use of detector confidence.

	Dynamic clutter	Distance-aware	Association type	Detection confidence	MOTA	HOTA
1	-	✓	Probabilistic	-	69.3	65.7
2	-	-	Probabilistic	-	71.7	65.7
3	-	-	IOU	-	73.7	65.4
3	✓	-	Probabilistic	-	75.6	66.4
4	✓	-	IOU	-	76.0	66.1
5	✓	✓	IOU	-	76.5	66.2
6	✓	✓	Probabilistic	-	77.1	68.2
BoT-SORT (Aharon et al., 2022)					78.5	69.2
7	✓	✓	Probabilistic	Raw	80.8	69.8
8	✓	✓	Probabilistic	Calibrated	81.6	70.2

The gray section of Tab. 1 reports results strictly using only the public detections of MOT17. Here, BASE outperforms vastly more complicated and computationally costly probabilistic trackers, even without visual odometry (VO).

5.6 MOT20 Results

The results for MOT20 are also shown in Fig. 1 and Tab. 1. Among submissions with publications and which use global parameters, BASE ranks first in MOTA, and third in HOTA. This is somewhat surprising, given that several of the other top methods employ Re-Id, which seems particularly promising on MOT20. The high HOTA-score might indicate that an empirically tuned motion model helps prevent mixing up targets during occlusions. On the testset, BASE (excl. the detector) ran at 39Hz on an AMD 5950x.

Several of the top-performing trackers employ per-sequence parameters, which seems to be particularly beneficial on MOT20. Sequences 04 and 07, the scenes of which are also featured in the training set, seem to warrant conservative tracker parameters. Meanwhile, sequences 06 and 08 seem to benefit from a more sensitive tracker.

5.7 DanceTrack

DanceTrack (Sun et al., 2022) is an interesting dataset as it poses quite different challenges than MOT17 and MOT20, with highly irregular motion but often relatively easily detectable targets. Since BASE’s motion model assumes continuous motion and slow changes, the sudden stretching of arms and changes in posture typical for this dataset seems particularly ill-suited for our model. Table 1 shows our results, as well as the SOTA methods also using the public detector. Our method outperforms ByteTrack (Zhang et al., 2021) and FineTrack (Ren et al., 2023) but falls behind C-BIoU, which leverages a more specialized motion model. OCSORT (Cao et al., 2022), which also uses a motion model adapted to DanceTrack, performs sim-

ilarly to BASE. MOTRv2 (Zhang et al., 2023) outperforms all these approaches by a large margin, illustrating that an end-to-end approach may be a particularly good fit for DanceTrack. However, as they use additional training data, the MOTRv2 results are not directly comparable to the other methods.

5.8 Ablation Study

In this section, we study the effects of the key components of BASE, namely the probabilistic association (vs IOU-based association), the distance-aware motion model (vs naive motion model), dynamic clutter estimation (vs constant λ_{EX}) and the histogram-calibrated detector confidence (vs ignoring or using raw confidence). We use the same YOLOX ablation model from (Zhang et al., 2021), so the ablation results are directly comparable to those of BoT-SORT (Aharon et al., 2022) and ByteTrack (Zhang et al., 2021). The model was trained on Crowdhuman (Shao et al., 2018) and the `train` half of the MOT17 training set. We fit all BASE-specific parameters using only `train` from MOT17, while the experiments were run on the `val` half of the training set. The results are shown in Tab. 2.

Comparing rows 1 and 2, we see that when using a constant λ_{EX} instead of dynamic clutter (Eq. (27)), the distance-aware motion model actually makes the method perform worse. Meanwhile, comparing rows 3 and 6, we see that the distance-aware motion model gives a significant boost once the dynamic clutter model is in place. Since the distance-aware motion model increases the position uncertainty for tracks with large bounding boxes, such tracks struggle to build confidence and match with detections when the clutter density is kept constant.

We also see that probabilistic association with a naive motion model (row 3) is outperformed by IOU-based association (row 4). The IOU-based association intrinsically offers some compensation for distance, since larger boxes are allowed to miss by more pixels while still achieving the same IOU as smaller boxes.

The poor result with the naive motion model (row 3) indicates that having some distance-aware mechanism is indeed necessary in visual tracking. Meanwhile, we also see that our proposed distance-aware motion model with probabilistic association (row 6) performs even better, indicating that BASE is able to further exploit the distance information.

We observe a significant improvement in performance when calibrated confidence is used (row 8) compared to raw confidence (row 7). Ignoring the confidence score altogether (row 6) results in worse performance than BoT-SORT. A possible explanation for this is that the BoT-SORT dual threshold approach can extract some, but not all, of the confidence score information. Since all proposed components are necessary to reach SOTA in the probabilistic paradigm, we consider BASE as a minimalist approach.

6 CONCLUSION

This paper demonstrates that a probabilistic tracker can achieve SOTA on popular VMOT benchmarks. Our proposed method, BASE, is the top-performing method on the MOT17 and MOT20 benchmarks and has competitive results on the more specialized DanceTrack benchmark. Through our ablation study, we show that a distance-aware motion model is necessary for probabilistic association to perform well in visual tracking, and that a dynamic clutter model is needed to make such motion models work. Previous attempts at probabilistic visual trackers omit (at least) the distance compensation, which we believe is why they fall behind more ad-hoc visual trackers that employ IOU-based association.

With BASE, we are merely scratching the surface of what is possible with probabilistic approaches to visual tracking. Starting from a minimalist probabilistic SHT foundation, we show that properly modeling motion, clutter, and detector confidence is all it takes for a probabilistic tracker to surpass the current SOTA. The probabilistic framework opens the gate for advanced core tracking algorithms and optimally exploiting multiple object features, such as visual appearance/Re-ID. Therefore, BASE is probably a better starting point for new visual trackers.

REFERENCES

- Aguilar, C., Ortner, M., and Zerubia, J. (2022). Small Object Detection and Tracking in Satellite Videos With Motion Informed-CNN and GM-PHD Filter. *Frontiers in Signal Processing*, 2(April):1–15.
- Aharon, N., Orfaig, R., and Bobrovsky, B.-Z. (2022). BoT-SORT: Robust Associations Multi-Pedestrian Tracking. *arXiv preprint arXiv:2206.14651*.
- Baisa, N. L. (2019). Online Multi-object Visual Tracking using a GM-PHD Filter with Deep Appearance Learning. *FUSION 2019 - 22nd International Conference on Information Fusion*.
- Baisa, N. L. (2021a). Occlusion-robust online multi-object visual tracking using a GM-PHD filter with CNN-based re-identification. *Journal of Visual Communication and Image Representation*, 80(January):103279.
- Baisa, N. L. (2021b). Robust online multi-target visual tracking using a HISP filter with discriminative deep appearance learning. *Journal of Visual Communication and Image Representation*, 77.
- Bar-Shalom, Y., Blackman, S. S., and Fitzgerald, R. J. (2007). Dimensionless score function for multiple hypothesis tracking. *IEEE Transactions on Aerospace and Electronic Systems*, 43(1):392–400.
- Bar-Shalom, Y. and Tse, E. (1975). Tracking in a cluttered environment with probabilistic data association. *Automatica*, 11(5).
- Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple Online and Realtime Tracking. *IEEE international conference on image processing (ICIP)*.
- Blackman, S. S. and Popoli, R. (1999). *Design and analysis of modern tracking systems*, volume 1999. Artech House Publishers.
- Bochinski, E., Eiselein, V., and Sikora, T. (2017). High-Speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017*.
- Brekke, E. F. (2019). *Fundamentals of Sensor Fusion: Target tracking, Navigation and SLAM*. NTNU.
- Cao, J., Weng, X., Khirodkar, R., Pang, J., and Kitani, K. (2022). Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking. *arXiv preprint arXiv:2203.14360*.
- Cetintas, O., Brasó, G., and Leal-Taixé, L. (2023). Unifying Short and Long-Term Tracking With Graph Hierarchies.
- Dendorfer, P., Rezatofghi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., and Leal-Taixé, L. (2020). MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.
- Du, Y., Song, Y., Yang, B., and Zhao, Y. (2022). StrongSORT: Make DeepSORT Great Again. *arXiv*.
- Ess, A., Leibe, B., Schindler, K., and Van Gool, L. (2008). A mobile vision system for robust multi-person tracking. *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.
- Evangelidis, G. and Psarakis, E. (2008). Parametric Image Alignment Using Enhanced Correlation Coefficient Maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865.
- Fu, Z., Angelini, F., Chambers, J., and Naqvi, S. M. (2019). Multi-Level Cooperative Fusion of GM-PHD Filters for Online Multiple Human Tracking. *IEEE Transactions on Multimedia*, 21(9):2277–2291.

- Fu, Z., Feng, P., Angelini, F., Chambers, J., and Naqvi, S. M. (2018). Particle PHD Filter Based Multiple Human Tracking Using Online Group-Structured Dictionary Learning. *IEEE Access*, 6.
- Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). YOLOX: Exceeding YOLO Series in 2021. *arXiv*.
- Jinlong Yang, Peng Ni, Jiani Miao, and Hongwei Ge (2022). Improving visual multi-object tracking algorithm via integrating GM-PHD and correlation filters.
- Kim, C., Li, F., Ciptadi, A., and Rehg, J. M. (2015). Multiple Hypothesis Tracking Revisited. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4696–4704. IEEE.
- Leal-Taixé, L., Milan, A., Reid, I., Roth, S., and Schindler, K. (2015). MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. *arXiv*.
- Mahler, R. P. S. and Martin, L. (2003). Multitarget Bayes Filtering via First-Order Multitarget Moments. *IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS*, 39(4).
- Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K. (2016). MOT16: A Benchmark for Multi-Object Tracking. *arXiv*.
- Nasseri, M. H., Babaei, M., Moradi, H., and Hosseini, R. (2022). Fast Online and Relational Tracking. *arXiv*.
- Qin, Z., Zhou, S., Wang, L., Duan, J., Hua, G., and Tang, W. (2023). MotionTrack: Learning Robust Short-term and Long-term Motions for Multi-Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17939–17948.
- Reid, D. B. (1979). An Algorithm for Tracking Multiple Targets. *IEEE Transactions on Automatic Control*, 24(6):843–854.
- Ren, H., Han, S., Ding, H., Zhang, Z., Wang, H., and Wang, F. (2023). Focus On Details: Online Multi-object Tracking with Diverse Fine-grained Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11289–11298.
- Sanchez-Matilla, R., Cavallaro, A., and N, N. (2020). Motion Prediction for First-Person Vision Multi-object Tracking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12538 LNCS, pages 485–499. Springer Science and Business Media Deutschland GmbH.
- Sanchez-Matilla, R., Poiesi, F., and Cavallaro, A. (2016). Online multi-target tracking with strong and weak detections. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9914 LNCS:84–99.
- Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., and Sun, J. (2018). CrowdHuman: A Benchmark for Detecting Human in a Crowd. *arXiv*, pages 1–9.
- Song, Y.-M., Yoon, K., Yoon, Y.-C., Yow, K. C., and Jeon, M. (2019). Online Multi-Object Tracking With GM-PHD Filter and Occlusion Group Management. *IEEE Access*, 7:165103–165121.
- Stadler, D. and Beyerer, J. (2022). Modelling Ambiguous Assignments for Multi-Person Tracking in Crowds. *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2022*, pages 133–142.
- Stadler, D. and Beyerer, J. (2023). An Improved Association Pipeline for Multi-Person Tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3170–3179. IEEE.
- Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., and Luo, P. (2022). DanceTrack: Multi-Object Tracking in Uniform Appearance and Diverse Motion.
- Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., Wang, C., and Luo, P. (2020). TransTrack: Multiple Object Tracking with Transformer. *arXiv*.
- Wang, Y., Kitani, K., and Weng, X. (2020). Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. *arXiv*.
- Wojke, N., Bewley, A., and Paulus, D. (2018). Simple online and realtime tracking with a deep association metric. *Proceedings - International Conference on Image Processing, ICIP, 2017-Sept*:3645–3649.
- Wojke, N. and Paulus, D. (2017). Confidence-Aware probability hypothesis density filter for visual multi-object tracking. *VISIGRAPP 2017 - Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 6(Visigrapp):132–139.
- Yan, B., Jiang, Y., Sun, P., Wang, D., Yuan, Z., Luo, P., and Lu, H. (2022). Towards Grand Unification of Object Tracking. *arXiv*.
- Yang, F., Chang, X., Sakti, S., Wu, Y., and Nakamura, S. (2021). ReMOT: A model-agnostic refinement for multiple object tracking. *Image and Vision Computing*, 106:104091.
- Yang, F., Odashima, S., Masui, S., and Jiang Fujitsu Research, S. (2023). Hard To Track Objects With Irregular Motions and Similar Appearances? Make It Easier by Buffering the Matching Space. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4799–4808.
- Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., and Wei, Y. (2021). MOTR: End-to-End Multiple-Object Tracking with Transformer. *arXiv*.
- Zhang, S., Benenson, R., and Schiele, B. (2017). CityPersons: A diverse dataset for pedestrian detection. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua*:4457–4465.
- Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., and Wang, X. (2021). ByteTrack: Multi-Object Tracking by Associating Every Detection Box. *ECCV 2022, Proceedings*.
- Zhang, Y., Wang, T., and Zhang, X. (2023). MOTRv2: Bootstrapping End-to-End Multi-Object Tracking by Pretrained Object Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22056–22065.
- Zhu, T., Hiller, M., Ehsanpour, M., Ma, R., Drummond, T., Reid, I., and Rezatofighi, H. (2021). Looking Beyond Two Frames: End-to-End Multi-Object Tracking Using Spatial and Temporal Transformers. *arXiv*, pages 1–20.