BeliefTrack: A New Framework for Improving SORT-like Tracking Algorithms with Multi-feature Association and Confidence Management

Wei Xu, Dominique Gruyer, Sio-Song Ieng

BeliefTrack: A New Framework for Improving SORT-like Tracking Algorithms with Multi-feature Association and Confidence Management

Wei Xu[1] and Dominique Gruyer[2] and Sio-Song Ieng[3]

***Abstract*— DeepSORT [1], a widely recognized Kalman filter-based tracking-by-detection (KFTBD) algorithm, has inspired various derivative versions, named SORT-like algorithms in this paper. Building on our previous work and SORT-like algorithms, this paper introduces an innovative framework that integrates a multi-feature association based on belief theory, and several enhanced components based on confidence, such as the Kalman filter introducing associations confidence, which aims to improve tracking performance in various complex environments. The proposed method is designed to be generic and can be integrated into a SORT-like tracker for improvement. The performance and compatibility of the new framework, namely BeliefTrack, have been evaluated and validated by 1) applying it to several datasets containing various and complex environments, 2) comparing it to DeepSORT as the baseline and several variants versions from StrongSORT [2], and 3) adapting it to different detectors. In all cases, our BeliefTrack demonstrates improved results, sometimes significantly.**

***Index Terms*— MOT, SORT-like algorithm, Belief theory, Uncertainty, Ambiguity, Multi-feature fusion, Confidence management**

I. Introduction

Effectively perceiving the environment with robustness in all traffic situations is critical for enabling automated vehicles to operate in urban environments and on highways. Ensuring road user safety mandates comprehensive obstacle detection and continuous trajectory monitoring, containing both mobile and stationary obstructions. To address this challenge, various sensors such as cameras, LiDAR, and RADAR are currently embedded in vehicles or installed in intelligent and communicating infrastructure. Simultaneously, advanced algorithms are needed to process data in real time, detecting, classifying, and tracking all obstacles effectively. This challenge has been known for several decades under the name of Multi Objects Tracking (MOT).

The KFTBD-based trackers [1], [3], [4] are widely recognized in the context of MOT, which combines the Kalman filter [5] and the Hungarian algorithm, as Gruyer et al. have already proposed within the framework of MOT based on the theory of evidence [6] and extended it to multi-object and multi-hypothesis tracking in [7], to the best of our knowledge. Benefiting from the emergence of superior deep-learning-based detectors and feature extractors, by combining KFTBD-based SORT [3] type algorithms, some methods, such as DeepSORT, StrongSORT, and BoTSORT [8], have demonstrated powerful performance for MOT tasks. These methods can be used in various transport applications, such as on-board MOT for obstacle avoidance and path planning, the cooperative perception [9] and the monitoring of road and urban mobility, as evidenced by the IA city challenges [10], [11]. For instance, [12] proposed a solution to the 2021 Challenge Track 3: Multi-Camera Vehicle Tracking at City-Scale, focusing on crossroad zones where re-identification methods are also needed.

Many recent studies have proposed improvements based on DeepSORT, which are known as SORT-like algorithms. Compared to the simple CNN used in DeepSORT, newer approaches use more powerful feature extractors like Bag of Tricks (BoT) [13] and Multiple Granularity Network (MGN) [14], which notably improve performance but at the expense of increased computational demands. Another avenue for improvement involves improving the motion model, employing techniques such as Kalman filter with constant-velocity assumptions [4], Noise Scale Adaptive (NSA) Kalman filter [15], and Camera Motion Compensation (CMC) [2], [8], [4]. Recently, ConfTrack introduced a novel approach using Confidence Weighted Kalman-Update [16]. Regarding association, many SORT-like algorithms use both motion and appearance features for similarity measurement, similar to DeepSORT. For instance, StrongSORT combines these features via a weighted sum, while BoTSORT adopts a similar strategy but also integrates an IoU-ReID fusion pipeline to combine IoU distance. Furthermore, some novel matching strategies have emerged in recent work [4], [17].

While the studies mentioned above have shown promising results, challenges persist in enhancing SORT-like algorithms for MOT tasks. Complex scenarios such as occlusion, intersection, or motion blur introduce noisy features into the association process, leading to uncertainties and ambiguities that require resolution. Although using appearance as the main feature for the association has already demonstrated efficiency in different SORT-like algorithms, it can be sometimes sensitive and performance will fall in certain cases, such as very crowded scenarios. However, the introduction of additional features can potentially lead to conflicts and ambiguities. Therefore, it becomes crucial to achieve a robust association based on various features, as well as effectively manage conflicts and ambiguities.

In this paper, we propose an innovative framework for

*This work was supported by Horizon Europe AUGMENTED-CCAM project.

[1]Wei Xu is with COSYS, PICS-L, Université Gustave Eiffel, F-77454 Marne-la-Vallée, France / STIC, Université Paris-Saclay, 91190 Gif-Sur-Yvette, France wei.xu@univ-eiffel.fr

[2]Dominique Gruyer is with COSYS, PICS-L, Université Gustave Eiffel, F-77454 Marne-la-Vallée, France dominique.gruyer@univ-eiffel.fr

[3]Sio-Song Ieng is with COSYS, PICS-L, Université Gustave Eiffel, F-77454 Marne-la-Vallée, France sio-song.ieng@univ-eiffel.fr

improving the SORT-like algorithms, by integrating a multi-feature association based on belief theory and introducing different types of confidences into the enhanced components.

The main contributions can be summarized as follows:

- We propose a generic framework, that can be integrated into SORT-like algorithms, and improve their performance by performing association based on multi-feature fusion using belief theory. The integration process of BeliefTrack to the baseline is demonstrated.
- Several novel enhanced components based on the confidence (mass) of associations/detections/tracks are proposed to improve the performance of tracking.
- A comprehensive evaluation is performed on the performance and compatibility of the proposed framework.

The remainder of the paper is organized as follows. Section II outlines the theoretical foundation of the proposed framework. Section III details the framework itself. Section IV describes the experimental evaluation, with results and analysis provided in Section V. Finally, Section VI concludes the paper and discusses potential future work.

II. Theoretical Basis

In this section, we will introduce the theoretical foundation regarding applying the belief theory to BeliefTrack framework.

A. Belief Functions

The central concept of belief theory involves evaluating the confidence level in hypotheses by calculating "mass values". The key elements can be defined as follows:

1) Hypothesis: Regarding tracking applications, a hypothesis H can be defined as the definitive connection between detections and tracks, denoted as "Detection i is associated with a Track j", namely a potential association during the tracking. Each hypothesis has an associated mass value $m_{i,j}(H)$, which quantifies the likelihood or weights of that particular assignment being true, typically falling within the range of $[0, 1]$.

2) Frame of discernment (Ω): The perceptual detections obtained from the surrounding environment produce a lot of hypotheses to known tracks, each implying a potential association. Together, these hypotheses construct the frame of discernment Ω as in Equation 1.

$$\begin{aligned} \Omega &= \{H_1, H_2, \ldots, H_n\} \\ H_i \cap H_j &= \varnothing \quad \text{for} \quad \forall i \neq j \end{aligned} \tag{1}$$

3) Power Set (2^n): Within Ω, the power set is formed by combining single hypotheses and their conjunctions, to encompass all potential assignments as shown in Equation 2. Of particular note are two significant combinations: the disjunction of all single hypotheses creates the unknown hypotheses, representing uncertainty or ignorance. Conversely, the conjunction of all single hypotheses signifies the empty hypotheses, indicating that all hypotheses are invalid.

$$\begin{aligned} 2^n &= \{A \mid A \subseteq \Omega\} \\ &= \{\varnothing_\Omega, H_1, H_2, \ldots, H_n, H_1 \cup H_2, \ldots, \Omega\} \\ &\text{where} \quad \varnothing_\Omega = H_1 \cap H_2 \ldots \cap H_n \end{aligned} \tag{2}$$

4) Basic belief assignment (BBA): Each proposition A within the power set is assigned a mass $m^\Omega(A)$, representing the degree of belief. This mass indicates the proportion of all informative evidence supporting the assignments. The sum of masses belonging to the BBA equals one, as demonstrated in the Equation 3.

$$\begin{aligned} m^\Omega(\varnothing_\Omega) &= 0 \\ \sum_{A \in 2^\Omega} m^\Omega(A) &= 1 \end{aligned} \tag{3}$$

5) Hypothesis world extension: The construction of the power set operates on the premise of a closed hypothesis world. In this framework, the mass assigned to the empty element ($m^\Omega(\varnothing_\Omega)$) is consistently set to 0, as illustrated in Equation 3. In the context of tracking systems, this construction implies a scenario where all objects are either detected or all ground tracks are identified. However, such an ideal scenario may not always align with real-world conditions. For instance, in environments characterized by limited detection capabilities or occluded objects, the closed hypothesis world assumption may be violated. Consequently, the mass associated with the empty element may deviate from 0, indicating the presence of uncertainty or ambiguity regarding object detection or identification.

Building upon the Open World framework [18] introduced by Smets, Royere & Gruyer proposed the Extended Open World framework [19], which offers a clear definition of the mass on new hypotheses and conflicts between information sources. This framework represents a new hypothesis by introducing a new element labeled "*". In a tracking system, this denotes a scenario where a detection fails to associate with an existing track within the frame of discernment, or vice versa. Such associations with "*" are used to detect the presence of new tracks or the propagation of current tracks. Thereby, the frame of discernment Ω is expanded to include the new element "*", denoted Θ as in Equation 4.

$$\Theta = \{H_1, H_2, \ldots, H_n, *\} \tag{4}$$

B. Significance of each hypothesis

To enhance understanding and recall upon introducing the new framework in the subsequent section, it is essential to declare the definition of the mass discussed earlier. The summary of these masses can be expressed as follows:

- $m_i(\Theta)$ represents the degree of uncertainty, indicating that all hypotheses within Θ are considered valid and thus the information is indeterminate.
- $m_{i,j}(H)$ represents the degree to which a hypothesis is deemed true and can also be interpreted as a measure of its similarity between the detection and track.
- $m_{i,j}(\overline{H})$ represents the degree to which a hypothesis is deemed false and can also be interpreted as a measure of its dissimilarity between the detection and track.
- $m_i(H*)$ represents the degree to which the reference object can not be associated with any object in the frame of discernment, indicating that no hypothesis within the frame is valid.

- $m_i(\varnothing)$ represents the degree to which some sources of information are contradictory in the information that they provide, following the definition in Equation 2.

III. Framework

In this section, BeliefTrack framework is introduced. The backbone structure of DeepSORT is used as a baseline since is also a fundamental of various SORT-like algorithms. Figure 1 provides a visual depiction of how the proposed framework integrated into DeepSORT's backbone, showcasing its overarching architecture and the dynamic interactions among its key components. The main enhancement and modification focus on 3 aspects: Association pipeline, Kalman filter, and Tracks management.

A. Association Pipeline

The proposed association pipeline keeps the two-stage structure like most SORT-like algorithms typically comprise: In the first phase, the algorithm uses multiple features to establish associations between detections and confirmed tracks in consecutive frames. Subsequently, in the second phase, the algorithm performs a refinement step using Intersection over Union (IOU) based association, to establish associations between unmatched detections and tentative tracks. The detailed process is demonstrated in the Algorithm 1.

Algorithm 1 Association pipeline

Input:
  $\mathcal{T}$ ▷ Tracks set contains confirmed, tentative, and propagated tracks,
  $\mathcal{D}$ ▷ Detections set with confidence
Output: $\hat{\mathcal{T}}$ ▷ Updated tracks set
procedure Association($\mathcal{T}$, $\mathcal{D}$)
  /* first association */
  for $\mathcal{O} \in \{Track, Detection\}$ **do** ▷ Track or detection as observer
    $\mathcal{M}^{\mathcal{O}}_{initial} \leftarrow$ Generation$(\mathcal{T}_{confirmed}, \mathcal{D}, \mathcal{F})$ ▷ Generate initial mass set given on the feature set $\mathcal{F}$
    $\mathcal{M}^{\mathcal{O}}_{synthetic} \leftarrow$ Fusion$(\mathcal{M}^{\mathcal{O}}_{initial})$ ▷ Generate synthetic mass set given on initial mass set
    $\mathcal{M}^{\mathcal{O}}_{combine} \leftarrow$ Combine$(\mathcal{M}^{\mathcal{O}}_{synthetic})$ ▷ Generate combination mass set given on synthetic mass
  end for
  $\hat{\mathcal{T}} \leftarrow$ Assign(Global$(\mathcal{M}^{\mathcal{O}_T}_{combine}, \mathcal{M}^{\mathcal{O}_D}_{combine})$) ▷ Global mass matrix-based association (Hungarian algorithm)
  /* Second association */
  $\hat{\mathcal{T}} \leftarrow$ Assign(cost$(\hat{\mathcal{T}}_{tentative}, \mathcal{D}_{unmatched})$) ▷ IOU cost matrix-based association (Hungarian algorithm)
  return $\hat{\mathcal{T}}$ ▷ Return updated track set
end procedure

1) First Association: Different from using the fusion of cost matrix based on the appearance and motion distance as an association indicator [2], we proposed a more generic fusion based on the mass matrix of hypotheses between detections and tracks, by using three steps to extract as much information as possible from the track-to-detection (detection-to-track) similarity comparison.

Initial mass-set generation The initial step involves creating the initial mass sets by assessing the distance $d^f_{i,j}$ of detection i with track j across various features denoted as f. These features contain appearance, motion, and IOU, with corresponding distance metrics including cosine distance, Mahalanobis distance, and IOU overlap. Each initial mass set gives information about the hypotheses of association H_j, non-association $\overline{H_j}$, and the unknown Θ, denoted as $\{m^f_{i,j}(H_j), m^f_{i,j}(\overline{H_j}), m^f_{i,j}(\Theta)\}$. To differentiate between associated and non-associated masses, we utilize an acceptance threshold distance. If the distance exceeds this threshold, the associated mass is set to 0, focusing only on calculating non-associated masses. Conversely, if the distance falls below the threshold, the non-associated mass is set to 0, allowing for the exclusive calculation of associated mass. Further, the thresholds Δ^f_1 and Δ^f_2 are used to adjust the impartiality of association, with a common scenario being that Δ^f_1 and Δ^f_2 sharing the same value. In Equation 5, R_f denotes the reliability of distinct features we defined, and β is the non-associated mass increase factor.

$$
\begin{aligned}
m^f_{i,j}(H_j) &= \begin{cases} R_f * (1 - e^{-\frac{\Delta^f_1 - d^f_{i,j}}{\Delta^f_1 / 3}}) & \text{if } d^f_{i,j} \in [0, \Delta^f_1] \\ 0 & \text{if } d^f_{i,j} \in [\Delta^f_1, \infty] \end{cases} \\
m^f_{i,j}(\overline{H_j}) &= \begin{cases} 0 & \text{if } d^f_{i,j} \in [0, \Delta^f_2] \\ R_f * (1 - e^{-\frac{d^f_{i,j} - \Delta^f_2}{\beta}}) & \text{if } d^f_{i,j} \in [\Delta^f_2, \infty] \end{cases} \\
m^f_{i,j}(\Theta) &= \begin{cases} 1 - m^f_{i,j}(\overline{H_j}) & \text{if } d^f_{i,j} \in [0, \Delta^f_1] \\ 0 & \text{if } d^f_{i,j} \in [\Delta^f_1, \Delta^f_2] \\ 1 - m^f_{i,j}(H_j) & \text{if } d^f_{i,j} \in [\Delta^f_2, \infty] \end{cases}
\end{aligned}
\tag{5}
$$

Multi-feature fusion. After obtaining initial mass sets across various features, a synthetic mass set is calculated using Equation 6, and the mass set is denoted as $\{m_{i,j}(H_j), m_{i,j}(\overline{H_j}), m_{i,j}(\varnothing), m_{i,j}(\Theta)\}$. This step introduces the mass on conflict primarily due to similarity measurement fusion across different features, as well as the mass on new hypotheses.

$$
\begin{aligned}
m_{f_{1,..,n}}(H_j) &= \prod_{k=1}^{n}(1 - m_{f_k}(\overline{H_j})) - \prod_{k=1}^{n} m_{f_k}(\Theta) \\
m_{f_{1,..,n}}(\overline{H_j}) &= \prod_{k=1}^{n}(1 - m_{f_k}(H_j)) - \prod_{k=1}^{n} m_{f_k}(\Theta) \\
m_{f_{1,..,n}}(\Theta) &= \prod_{k=1}^{n} m_{f_k}(\Theta) \\
m_{f_{1,..,n}}(\varnothing) &= 1 - m_{f_{1,..,n}}(H_j) - m_{f_{1,..,n}}(\overline{H_j}) \\
&\quad - m_{f_{1,..,n}}(\Theta)
\end{aligned}
\tag{6}
$$

BBA combination. The goal of the BBA combination is to enhance the reliability of associations by combining all information about observed objects from

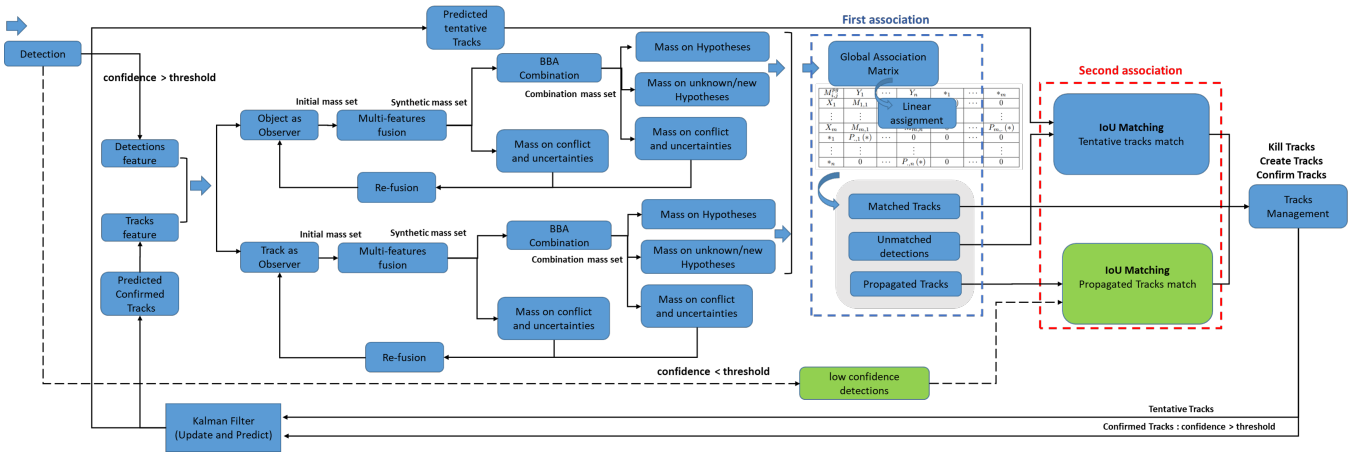

Fig. 1. Architecture of BeliefTrack framework, integrated into DeepSORT

their observer, which could be a detection or a track. For each observed object, the synthetic mass set is built independently from the other observed objects, even if they are related to the same observer. Therefore, different information pieces (namely the mass from the synthetic mass set) will be combined in this step from other observed objects, to regenerate the mass related to the degree of association and non-association ($m_{i,.}(H_j)$ and $m_{i,.}(\overline{H_j})$) for each potential association. The mass of uncertainty, conflict, unknown ($m_{i,.}(H_*), m_{i,.}(\Theta), m_{i,.}(\varnothing)$) will be generated for each observer. The mass of the unknown (new) hypothesis (H_*) within the extended open-world framework is introduced in this step. The set of equations for calculating the new mass sets is presented in Equation 7.

$$
\begin{aligned}
m_{i,.}(H_j) &= m_{i,j}(H_j) \prod_{\substack{\omega=1..k \\ \omega \neq j}} (1 - m_{i,\omega}(H_\omega)) \\
m_{i,.}(H_*) &= \prod_{j=1..k} m_{i,j}(\overline{H_j}) \\
m_{i,.}(\Theta) &= \prod_{j=1..k} (m_{i,j}(\Theta) + m_{i,j}(\overline{H_j})) - \prod_{j=1..k} m_{i,j}(\overline{H_j}) \\
m_{i,.}(\varnothing) &= 1 - \prod_{j=1..k} (m_{i,j}(\Theta) + m_{i,j}(\overline{H_\omega})) \\
&\quad - \sum_{j=1..k} m_{i,j}(H_j) \prod_{\substack{\omega=1..k \\ \omega \neq j}} (1 - m_{i,\omega}(H_\omega))
\end{aligned}
\tag{7}
$$

In this step, the combination mass set is generated from the view of the different observers (tracks or detections), which is denoted as $\{m_{i,.}(H_j), m_{i,.}(\overline{H_j}), m_{i,.}(H_*), m_{i,.}(\Theta), m_{i,.}(\varnothing)\}$.

Association based on global matrix. The mass sets derived from BBA combination contain all the mass information concerning the detection or tracking as observed by observers ($m_{i,.}$ and $m_{.,j}$). The first observer assesses the potential association degree between the observed tracks $\mathcal{T}_{j\in[0,n]}$ and known detection $\mathcal{D}_i$, represented by mass $m_{i,.}(H_j)$. Here, the unknown hypothesis H_* signifies no track being associated with the current detection $\mathcal{D}_i$. Conversely, the second observer evaluates the potential association degree between the observed detection $\mathcal{D}_{i\in[0,m]}$ and known track $\mathcal{T}_j$, with mass $m_{.,j}(H_i)$, where H_* denotes no detection being associated with the current track $\mathcal{T}_j$. The fusion of these two mass sets contributes to the final mass matrix (global association matrix), as illustrated in Equation 8.

$$
\begin{aligned}
\mathcal{P}_{i,j}(H_{i,j}) &= m_{i,.}(H_j) \times m_{.,j}(H_i) \\
\mathbf{Diag}(\mathcal{P}_{i\in[1,m],j\in[n,n+m]}(H_*)) &= m_{i,.}(H_*) \\
\mathbf{Diag}(\mathcal{P}_{i\in[m,m+n],j\in[1,n]}(H_*)) &= m_{.,j}(H_*)
\end{aligned}
\tag{8}
$$

The format of the global association matrix is clearly illustrated in Table I.

TABLE I
GLOBAL ASSOCIATION MATRIX FORMAT

$\mathcal{P}_{i,j}(H_{i,j})$	$\mathcal{T}_1$	…	$\mathcal{T}_n$	$*_1$	…	$*_m$
$\mathcal{D}_1$	$\mathcal{P}_{1,1}(H_{1,1})$	…	$\mathcal{P}_{1,n}(H_{1,n})$	$m_{1,.}(H_*)$	…(0)	0
⋮	⋮	⋱	⋮	⋮(0)	⋱	⋮(0)
$\mathcal{D}_m$	$\mathcal{P}_{m,1}(H_{m,1})$	…	$\mathcal{P}_{m,n}(H_{m,n})$	0	…(0)	$m_{m,.}(H_*)$
$*_1$	$m_{.,1}(H_*)$	…(0)	0	0	…(0)	0
⋮	⋮(0)	⋱	⋮(0)	⋮(0)	⋱(0)	⋮(0)
$*_n$	0	…(0)	$m_{.,n}(H_*)$	0	…(0)	0

In the final step, we execute the linear assignment process based on the global association matrix. Since the masses within a track's or detection's frame of discernment are strictly within the range of 0 to 1, sometimes even approach values to 0, a normalization step is performed necessarily. Furthermore, to transform the global mass matrix into the cost matrix, we subtract the normalized values from 1, following the computation detailed in Equation 9.

$$
C_{i,j}(H_{i,j}) = 1 - \|\mathcal{P}_{i,j}(H_{i,j})\| \tag{9}
$$

2) Second Association: The second association step involves IOU matching for unmatched detections and tentative tracks, a common practice in SORT-like algorithms. This

process aims to enhance the tracking by further discovering potential matches based on the overlap between bounding boxes. Furthermore, an emerging trend is to perform an association step between low-confidence detections with propagated tracks (confirmed but losing matched detection). T This is reasonable because low-confidence detections often result from occlusion or environmental factors that can also cause confirmed tracks to lose matches. This approach has demonstrated efficiency in [4], [8].

B. Second-stage Fusion (SsF)

In the multi-feature fusion and BBA combination steps, we obtain the mass of uncertainty and conflict. The degree of uncertainty comes from the mass of the union of all of the hypotheses, namely $m(\Theta)$, the high mass value indicating inefficiency in distinguishing potential associations by our initial feature sets, which comprise appearance and motion (A, M). Introducing a new feature like IOU (I) becomes relevant in such a situation. Conflict masses $m(\varnothing)$ stem mainly from similarity conflicts across features. Typically, appearance features start with higher reliability. Adjusting reliability weights between selected feature sets, such as boosting motion feature reliability while reducing appearance feature reliability, will be relevant in optimizing fusion settings. After adjusting of feature set or reliability, the first two steps of mass calculations in the first association will be re-performed, and the synthetic mass set will be updated accordingly and sent to the next step. The second-stage fusion process integrated into the framework is illustrated by the pseudo-code provided in Algorithm 2.

Algorithm 2 SECOND-STAGE FUSION

Input:
  $\mathcal{M}_{syn}$ or $\mathcal{M}_{com}$ ▷ Original synthetic mass set and BBA combination mass set
  $\Delta_\Theta, \Delta_\varnothing$ ▷ Thresholds of uncertainty and conflict mass
Output: $\hat{\mathcal{M}}_{syn}$ ▷ Updated synthetic mass set
procedure 2ND_FUSION(**Input**)
  if $\mathcal{M}^i_{syn}(\Theta) > \Delta_\Theta$ or $\mathcal{M}^i_{com}(\Theta) > \Delta_\Theta$ **then**
    $\mathcal{F}_{\{A,M,I\}} \leftarrow \mathcal{F}_{\{A,M\}}$ ▷ Add IOU as feature into initial feature sets
    $\hat{\mathcal{M}}_{syn} \leftarrow$ REFUSION$(\mathcal{F}_{\{A,M,I\}}, i)$ ▷ Regeneration initial mass set for Object i, and then process multi-feature fusion step to update synthetic mass set
  end if
  if $\mathcal{M}^i_{syn}(\varnothing) > \Delta_\varnothing$ or $\mathcal{M}^i_{com}(\varnothing) > \Delta_\varnothing$ **then**
    $\{\mathcal{R}_A \downarrow, \mathcal{R}_M \uparrow\} \leftarrow$ MODIFY$(\{\mathcal{R}_A, \mathcal{R}_M\})$ ▷ Modify the reliability of appearance and motion feature
    $\hat{\mathcal{M}}_{syn} \leftarrow$ REFUSION$(\{\mathcal{R}_A \downarrow, \mathcal{R}_M \uparrow\}, i)$ ▷ Regeneration initial mass set for Object i, and process multi-feature fusion to update synthetic mass set
  end if
  return $\hat{\mathcal{M}}_{syn}$ ▷ Return updated synthetic mass set
end procedure

C. Confidence Fused Mass Generation (CFMG)

Inspired by [4], [16], where the cost matrix is fused with confidence, aiming for robust tracking operations in the ID matching process. In our framework, the confidence of Detection $C_\mathcal{D}$ is introduced during mass generation. Unlike previous approaches, we employ complementary confidence for appearance, motion, and IOU features, using $C_\mathcal{D}$ for appearance and $1 - C_\mathcal{D}$ for motion and IOU. Equation 10 illustrates the fusion of confidence in generating the mass of hypotheses of association, the process for hypotheses of non-association is in the same manner.

$$\begin{aligned} m^A_{i,j}(H_j) &= C_{\mathcal{D}_i} \times R_A \times (1 - e^{-\frac{\Delta^A_1 - d^A_{i,j}}{\Delta^A_1/3}}) \\ m^M_{i,j}(H_j) &= (1 - C_{\mathcal{D}_i}) \times R_M \times (1 - e^{-\frac{\Delta^M_1 - d^M_{i,j}}{\Delta^M_1/3}}) \\ m^I_{i,j}(H_j) &= (1 - C_{\mathcal{D}_i}) \times R_I \times (1 - e^{-\frac{\Delta^I_1 - d^I_{i,j}}{\Delta^I_1/3}}) \end{aligned} \tag{10}$$

D. Mass Fused Kalman Filter (MFKF)

The Kalman filter is a key component of the SORT-like algorithm for multi-object tracking, which operates recursively, with two main steps: prediction and update.

In the update step, the Kalman filter corrects the predicted state based on the observed measurements (matched detection), where we introduced mass of the hypothesis of association as weights. The weighted updates of state vector and estimate covariance are given by Equation 11:

$$\hat{x}_{j_k|k} = \begin{cases} \hat{x}_{j_k|k-1} & + \ m_{i,j_k} * K_{i,j_k}(z_{i_k} - H_{j_k}\hat{x}_{j_k|k-1}) \\ & \text{if } m_{i,j_k} \in [\Delta^m_1, \Delta^m_2] \\ \hat{x}_{j_k|k-1} & + \ K_{i,j_k}(z_{i_k} - H_{j_k}\hat{x}_{j_k|k-1}) \\ & \text{else} \end{cases}$$

$$P_{j_k|k} = (I - K_{i,j_k}H_{j_k})P_{j_k|k-1} + (m_{i,j_k})^2 * K_{i,j_k}R_{k_i}K^T_{i,j_k} \tag{11}$$

where K_{i,j_k} is the Kalman gain, H_{j_k} is the measurement matrix, z_{i_k} is the measurement vector, and R_{k_i} is the measurement noise covariance matrix. $\hat{x}_{j_k|k}$ is the updated state estimate, and $P_{j_k|k}$ is the updated covariance matrix. Note that the state vector undergoes updates only within the low mass score range, represented by $[\Delta^m_1, \Delta^m_2]$.

These steps are iterated for each time step k, resulting in an accurate estimation of the object states over time.

E. Trust Fused Track Management (TFTM)

The maintenance of the track is a crucial aspect of tracking, and it is responsible for creating, confirming, propagating, and deleting. We introduce the trust estimation of a track inside our framework, which contributes to the management of the track. The estimation process is demonstrated as in Equation 12.

$$\begin{aligned} \eta &= -\tau \times \ln(1 - C^{k-1}_{\mathcal{T}_j}) \\ C^k_{\mathcal{T}_j} &= 1 - e^{-\eta/\tau} \\ &\text{if } \mathcal{T}_j \text{ is associated } \ \eta{+}{+} \ \text{ else } \ \eta{-}{-} \end{aligned} \tag{12}$$

When a track $\mathcal{T}_j$ is associated with a certain detection, the parameter η increases, thereby enhancing the trust value $C^k_{\mathcal{T}_j}$.

Conversely, when $\mathcal{T}_j$ is no longer associated, η decreases, leading to a reduction in $C^k_{\mathcal{T}_j}$. Furthermore, τ serves as the adjustment factor in the exponential trust variation function.

The management of tracks based on trust can be summarized as:

- The tentative track $\mathcal{T}_j$ is confirmed if the trust value $C^k_{\mathcal{T}_j}$ exceeds the threshold $\Delta_{confirm}$.
- The tentative track $\mathcal{T}_j$ is deleted if the trust value $C^k_{\mathcal{T}_j}$ falls below the threshold Δ_{delete}.
- The propagated track $\mathcal{T}_j$ is deleted if its lifespan (number of frames present) $\mathcal{L}_{\mathcal{T}_j}$ exceeds the threshold $\Delta_{lifespan}$.

IV. Experiments

A. Experiments Configuration

1) Public Datasets: Following common practices in tracking algorithm evaluation, we employ two widely recognized datasets: the MOT17 dataset [20] and the more recent MOT20 dataset [21]. The former offers complex scenarios for multi-person tracking, encompassing diverse challenges such as occlusions, camera motion, and various lighting conditions, the latter presents even more challenging scenarios characterized by high density. We also create MOT17 half-validation and half-training sets by partitioning the MOT17 training set, following the approach outlined by previous researchers [4]. Moreover, we adopt the KITTI left color image dataset [22], which is related to autonomous driving tasks, enabling tracking of pedestrians and vehicles.

2) Evaluation Metrics: We use TrackEval [23] as our evaluation standard to calculate three main categories of metrics: HOTA [24], CLEAR [25], and Identity [26]. HOTA serves as our primary metric, and we report it alongside association-related metrics such as Accuracy (AssA), Recall (AssRe), and Precision (AssPr). Additionally, we include widely adopted CLEAR metrics like multi-object tracking accuracy (MOTA) and multi-object prediction (MOTP). To comprehensively evaluate tracking continuity and re-identification accuracy, we also incorporate IDF1.

B. Implementation Basis

1) Detector: Public detections from MOT17 (FRCNN [27], DPM [28], and SDP [29]) and MOT20 serve as our first category for comparison with our baseline. In alignment with previous research, we employ YOLOX [30] as our detector on the MOT17 half-validation set for ablation studies, and also for evaluation of our framework's compatibility with other SORT-like algorithms. We adopt the weights of the ablation model from ByteTrack [4], which is trained on the CrowdHuman [31] and MOT17 half-training set. To further assess our framework's compatibility with multiple detectors, we utilize YOLOX, YOLOv5 [32], and YOLOv8 [33], all accompanied by publicly available weights trained on the COCO dataset [34].

2) Hyperparameters: The hyperparameters we used for the framework during experiments are follows:

- For primary mass set generation, reliability sets contains R_A: 1.0 for BoT, 0.4 for simple CNN, R_{motion}: 0.1, R_{IOU}: 0.01; Threshold sets contains Δ_A: 0.35 for BoT, 0.15 for simple CNN, Δ_{motion}: 2.0, Δ_{IOU}: 0.1, β: 0.3.
- For SsF, both threshold of conflict $\Delta_{\varnothing}$ and uncertainties Δ_{Θ} are 0.3, new reliability set contains R^{new}_A: 0.95 for BoT, 0.35 for simple CNN, R^{new}_{motion}: 0.5.
- For MFKF, the low mass score range $[\Delta^m_1, \Delta^m_2]$ are [0.2, 0.5].
- For TFTM, τ is 5.0, threshold $\Delta_{confirm}$ is 0.8, threshold Δ_{delete} is 0.3, threshold $\Delta_{lifespan}$ is 30.
- All experiments are equipped with a 16-core Intel i9-12950HX 2.30 GHz CPU and NVIDIA-GTX A4500 GPU core graphics card.

V. Results

In this section, we will showcase multiple results to provide a thorough evaluation of the performance and compatibility of the BeliefTrack framework. We employed the reproduced version of DeepSORT* from StrongSORT as our baseline, which also serves as the foundation for integrating our framework.

A. Comparison to baseline

The initial test focuses on comparing tracking outcomes based on the public detections provided by MOT17 and MOT20. We chose to implement two components SsF and TFTM in BeliefTrack due to the lack of confidence representation in the public detections. As shown in Table II, the improvement of BeliefTrack over the baseline is apparent, with notable advancements in HOTA, ASSA, and IDF1 metrics. This demonstrates that integrating our framework provides a stronger association and consistency of ID matching during tracking under the same conditions. In MOT20, accompanying high-density and complex scenarios introduce more conflict and ambiguity, this kind of increase in performance is even more pronounced.

TABLE II
Comparison between BeliefTrack and baseline, based on the public detections provided by MOT17 and MOT20

Dataset	Tracker	HOTA	AssA	IDF1	MOTA	MOTP
MOT17-FRCNN	Baseline	44.02	43.54	49.63	49.10	86.48
	Ours	47.42	50.16	55.16	49.22	86.45
MOT17-DPM	Baseline	28.22	29.18	34.72	29.97	77.76
	Ours	30.45	33.37	37.92	29.72	77.70
MOT17-SDP	Baseline	51.00	46.87	58.78	64.87	84.39
	Ours	54.74	53.67	64.43	65.08	84.51
MOT20	Baseline	38.55	31.53	42.95	53.80	86.11
	Ours	42.87	38.93	50.70	54.73	86.29

B. Compatibility to other SORT-like algorithm

DeepSORT continues to evolve with the integration of increasingly advanced techniques, as evidenced by the presentation of various advanced modules in StrongSORT and the proposal of different variants. One of the objectives of BeliefTrack is to introduce a generic framework for SORT-like algorithms. Therefore, it is crucial to assess the

TABLE III

COMPARISON BETWEEN BELIEFTRACK AND STRONGSORT VERSIONS ON MOT17 VALIDATION SET

Tracker	BoT	CMC	NSA	EMA	MC	woC	AFlink	GSI	HOTA	AssA	AssRe	AssPr	IDF1	IDR	IDP	MOTA	MOTP
Baseline	-	-	-	-	-	-	-	-	66.26	67.34	73.41	79.64	77.33	72.02	83.49	76.71	84.58
Ours	-	-	-	-	-	-	-	-	67.17	68.89	74.38	80.72	78.55	73.84	83.90	76.87	84.60
StrongSORTv1	✓	-	-	-	-	-	-	-	67.83	70.45	74.98	84.73	79.49	73.82	86.10	76.77	84.79
Ours	✓	-	-	-	-	-	-	-	68.42	71.33	76.47	84.69	80.43	75.56	86.00	77.05	84.75
StrongSORTv3	✓	✓	✓	-	-	-	-	-	68.27	70.82	75.31	85.21	79.67	74.07	86.20	77.12	85.11
Ours	✓	✓	✓	-	-	-	-	-	68.60	71.59	76.43	85.13	80.75	75.91	86.24	77.25	84.68
StrongSORTv6	✓	✓	✓	✓	✓	✓	-	-	68.90	72.12	76.86	85.72	81.45	75.70	88.15	76.91	85.22
Ours	✓	✓	✓	✓	-	-	-	-	70.13	74.65	78.57	86.45	83.66	78.71	89.29	76.83	84.98
StrongSORT++	✓	✓	✓	✓	✓	✓	✓	✓	70.18	73.36	78.57	85.38	82.47	78.74	86.57	78.60	85.02
Ours	✓	✓	✓	✓	-	-	✓	✓	71.49	76.14	80.86	85.64	84.78	81.76	88.02	78.63	84.81

compatibility of our framework with these variants. Table III shows the results compared to different versions of StrongSORT, which integrates the combination of several modules: BoT (S50) model from FastReID [35], Camera Motion Compensation (CMC), Noise Scale Adaptive (NSA) Kalman filter, Exponential Moving Average (EMA) feature updating mechanism, matching with Motion Cost (MC), and abandoning matching Cascade (woC), Appearance-Free link model (AFLink), Gaussian-Smoothed Interpolation (GSI).

The results demonstrate that our framework can enhance various variants of StrongSORT, as evidenced by a significant increase in the metric of the HOTA category and Identity category. However, Table II and Table III show our framework does not have a large margin on MOTA and MOTP compared to HOTA, AssA, and IDF1, but still demonstrates improvement in some cases.

C. Ablation studies

We conducted an ablation study on the MOT17-val dataset, using StrongSORTv6 as the foundation for integrating our framework. The experiments were initially performed with each enhanced component (CFMG, MKMF, TFTM, SsF) proposed in BeliefTrack individually, followed by their combination. The results in Table IV show consistent contributions from CFMG, TFTM, and SsF to our framework, e.g., with an improvement of around 0.2 in HOTA. However, a separate application of MFKF was found to have a negative influence on the metric. It is because, without the detection confidence introduced into the mass generation by CFMG, the mass is not suited for weighting target measurement. The combination of CFMG and MKMF demonstrates improvement, as shown in Table IV. Furthermore, introducing SsF and TMTF based on this combination apparently enhances the association aspect.

D. Compatibility to other detectors

The compatibility of the proposed BeliefTrack with detections from other detectors is demonstrated in Table V. We use the lightweight versions of three YOLO detectors, with online weights (trained on COCO dataset). Across all detectors, the integration of BeliefTrack led to improved metric values compared to the baseline.

TABLE IV

ABLATION STUDIES ON MOT17 VALIDATION SET

CFMG	MFKF	TFTM	SsF	HOTA	AssA	IDF1	MOTA	MOTP
-	-	-	-	69.30	73.06	82.41	77.00	85.04
✓	-	-	-	69.40	73.10	82.32	77.04	85.10
-	✓	-	-	69.23	73.00	82.34	76.89	85.02
-	-	✓	-	69.53	73.25	82.48	77.02	85.00
-	-	-	✓	69.38	73.28	82.60	76.96	85.02
✓	✓	-	-	69.42	73.36	82.52	76.88	85.07
✓	✓	✓	-	69.78	73.91	82.86	76.88	85.00
✓	✓	✓	✓	70.13	74.65	83.66	76.83	84.98

TABLE V

COMPARISON BETWEEN BELIEFTRACK AND BASELINE, USING 3 DIFFERENT DETECTORS ON KITTI COLOR IMAGES SET

Vehicle						
Detector	**Tracker**	**HOTA**	**AssA**	**IDF1**	**MOTA**	**MOTP**
YOLOXs	Baseline	48.54	56.30	64.01	49.33	74.58
	Ours	49.81	58.00	65.56	49.40	74.23
YOLOv5s	Baseline	48.66	57.70	63.77	49.02	78.00
	Ours	49.64	57.90	64.48	50.13	77.92
YOLOv8s	Baseline	49.90	57.71	65.60	52.53	77.25
	Ours	50.87	58.09	66.23	53.50	77.06

Pedestrian						
Detector	**Tracker**	**HOTA**	**AssA**	**IDF1**	**MOTA**	**MOTP**
YOLOXs	Baseline	38.18	42.94	56.66	35.28	69.70
	Ours	39.74	45.48	59.04	35.00	69.47
YOLOv5s	Baseline	38.86	43.18	56.92	38.39	71.46
	Ours	40.36	45.33	59.12	38.92	71.30
YOLOv8s	Baseline	39.87	43.81	58.00	38.71	71.52
	Ours	41.71	46.98	60.96	38.85	71.44

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduce BeliefTrack, a novel framework designed to enhance the SORT-like algorithm by performing the association phase based on multi-feature fusion using belief theory. Furthermore, several novel enhanced techniques are integrated based on the management of confidence of association, detection, and tracks, such as SsF, CFMG, MFKF, and TFTM to improve the performance of the tracking algorithm further, especially in association and

ID matching aspects. Through evaluation and validation in comprehensive experiments, the framework is demonstrated to effectively enhance tracking performance compared to the baseline, sometimes significantly. Furthermore, it is validated to be compatible with various advanced techniques that are applied to the SORT-like algorithm, as well as various detectors. These results support our claims that BeliefTrack is designed in generic to be integrated into SORT-like tracking algorithms to improve tracking performance.

In future work, attention could be directed toward several aspects: 1) Extend the framework into the integration of different and complementary sensor technologies such as LiDAR, RADAR, IR cameras, and neuromorphic cameras. 2) We'll also investigate the framework's capability to dynamically consider the quality of extractors, adjusting the reliability of data sources and features accordingly. 3) BeliefTrack will be explored with more advanced techniques or modules to show its compatibility.

REFERENCES

[1] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[2] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng, "Strongsort: Make deepsort great again," *IEEE Transactions on Multimedia*, 2023.

[3] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[4] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *European conference on computer vision*. Springer, 2022, pp. 1–21.

[5] R. G. Brown and P. Y. Hwang, "Introduction to random signals and applied kalman filtering: with matlab exercises and solutions," *Introduction to random signals and applied Kalman filtering: with MATLAB exercises and solutions*, 1997.

[6] D. Gruyer and V. Berge-Cherfaoui, "Multi-objects association in perception of dynamical situation." in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence(UAI'99)*, 1999.

[7] D. Gruyer, S. Demmel, V. Magnier, and R. Belaroussi, "Multi-hypotheses tracking using the dempster–shafer theory, application to ambiguous road context," *Information Fusion*, vol. 29, pp. 40–56, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253515000962

[8] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-sort: Robust associations multi-pedestrian tracking," *arXiv preprint arXiv:2206.14651*, 2022.

[9] S.-S. Ieng, M. Paget, M. A. Tchalim, and J.-P. Tarel, "Road side perception systems for safer intersections: Field test," in *2023 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, 2023, pp. 293–299.

[10] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, X. Yang, Y. Yao, L. Zheng, P. Chakraborty, C. E. Lopez, A. Sharma, Q. Feng, V. Ablavsky, and S. Sclaroff, "The 5th ai city challenge," 2021.

[11] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, Y. Yao, L. Zheng, M. S. Rahman, A. Venkatachalapathy, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, A. Li, S. Li, and R. Chellappa, "The 6th ai city challenge," 2022.

[12] C. Liu, Y. Zhang, H. Luo, J. Tang, W. Chen, X. Xu, F. Wang, H. Li, and Y.-D. Shen, "City-scale multi-camera vehicle tracking guided by crossroad zones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4129–4137.

[13] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[14] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[15] Y. Du, J. Wan, Y. Zhao, B. Zhang, Z. Tong, and J. Dong, "Giaotracker: A comprehensive framework for mcmot with global information and optimizing strategies in visdrone 2021," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 2809–2819.

[16] H. Jung, S. Kang, T. Kim, and H. Kim, "Conftrack: Kalman filter-based multi-person tracking by utilizing confidence score of detection box," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6583–6592.

[17] D. Stadler and J. Beyerer, "An improved association pipeline for multi-person tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3170–3179.

[18] P. Smets, "Non-standard logics for automated reasoning," 1988.

[19] C. Royère, D. Gruyer, and V. Cherfaoui, "Data association with believe theory," in *Proceedings of the Third International Conference on Information Fusion*, vol. 1. IEEE, 2000, pp. TUD2–3.

[20] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[21] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "Mot20: A benchmark for multi object tracking in crowded scenes," *arXiv preprint arXiv:2003.09003*, 2020.

[22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[23] A. H. Jonathon Luiten, "Trackeval," https://github.com/JonathonLuiten/TrackEval, 2020.

[24] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "Hota: A higher order metric for evaluating multi-object tracking," *International journal of computer vision*, vol. 129, pp. 548–578, 2021.

[25] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.

[26] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European conference on computer vision*. Springer, 2016, pp. 17–35.

[27] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[28] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[29] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2129–2137.

[30] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[31] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *arXiv preprint arXiv:1805.00123*, 2018.

[32] G. Jocher, "YOLOv5 by Ultralytics," May 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[33] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[35] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "Fastreid: A pytorch toolbox for general instance re-identification," *arXiv preprint arXiv:2006.02631*, 2020.