mathematics

MDPI

Article

Multi-Frame Joint Detection Approach for Foreign Object Detection in Large-Volume Parenterals

Ziqi Li, Dongyao Jia *, Zihao He and Nengkai Wu

School of Automation and Intelligence, Beijing Jiaotong University, Beijing 100044, China; 21111057@bjtu.edu.cn (Z.L.); 22115032@bjtu.edu.cn (Z.H.); 19111054@bjtu.edu.cn (N.W.)
* Correspondence: dongyaojia1974@163.com

Abstract: Large-volume parenterals (LVPs), as essential medical products, are widely used in healthcare settings, making their safety inspection crucial. Current methods for detecting foreign particles in LVP solutions through image analysis primarily rely on single-frame detection or simple temporal smoothing strategies, which fail to effectively utilize spatiotemporal correlations across multiple frames. Factors such as occlusion, motion blur, and refractive distortion can significantly impact detection accuracy. To address these challenges, this paper proposes a multi-frame object detection framework based on spatiotemporal collaborative learning, incorporating three key innovations: a YOLO network optimized with deformable convolution, a differentiable cross-frame association module, and an uncertainty-aware feature fusion and re-identification module. Experimental results demonstrate that our method achieves a 97% detection rate for contaminated LVP solutions on the LVPD dataset. Furthermore, the proposed method enables end-to-end training and processes five bottles per second, meeting the requirements for real-time pipeline applications.

Keywords: object detection; multi-object track; cross-frame feature association; large-volume parenterals; re-identification

MSC: 68T10; 68T20; 68T45

check for updates

Academic Editors: Xujuan Zhou, Yuefeng Li, Raj Gururajan, Ji Zhang and Revathi Venkataraman

Received: 29 March 2025
Revised: 15 April 2025
Accepted: 16 April 2025
Published: 18 April 2025

Citation: Li, Z.; Jia, D.; He, Z.; Wu, N. Multi-Frame Joint Detection Approach for Foreign Object Detection in Large-Volume Parenterals. *Mathematics* **2025**, *13*, 1333. https://doi.org/10.3390/math13081333

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Large-volume parenterals (LVPs), commonly known as infusion solutions, are sterile liquid preparations with volumes greater than or equal to 50 mL that are administered intravenously [1]. As one of the five major pharmaceutical formulations in China (tablets, capsules, injections, lyophilized powder injections, and infusion solutions), LVP solutions represent the most widely used therapeutic approach in clinical applications. In 2023, China's LVP industry production reached 10.6 billion bottles/bags. With ongoing healthcare reforms and increasing health awareness among residents, the social demand for LVP solutions continues to grow. However, due to manufacturing process limitations, various foreign particles such as hair, insects, plastic debris, and short fibers may contaminate LVP solutions during production. If these contaminants enter the human body undetected, they pose significant risks to patient safety. Therefore, LVP safety is a crucial aspect of pharmaceutical safety, making its improvement particularly significant.

Currently, most factories still rely on manual light inspection methods for LVP quality control [2]. This process involves workers examining bottles under specific lighting conditions, manually shaking or rotating the products to detect visible foreign particles such as

hair or plastic debris. While effective, manual inspection has notable limitations: inconsistent detection quality due to inspector fatigue, potential health risks to workers from prolonged exposure to strong lighting, insufficient inspection speed despite increased labor force, and possible secondary contamination during handling. Consequently, transitioning to automated inspection equipment represents an inevitable trend for improving detection efficiency and ensuring LVP safety.

Recent advances [3] in deep learning-based image recognition technology have provided new approaches to foreign particle detection in LVP solutions. Our research has successfully addressed challenges such as small object detection and interference from similar objects like bubbles and scratches, proposing a novel detection framework with three key innovations:

1. A YOLOV9S-based detection network optimized with Deformable Convolutional Networks (DCNs) [4], enhancing detection capabilities for various foreign particles and improving adaptation to refractive distortions.
2. A differentiable multi-object tracking mechanism that incorporates motion prediction, spatiotemporal correlation modeling, Sinkhorn algorithm [5], and uncertainty-based trajectory management, achieving end-to-end trainable object tracking with improved robustness and adaptability.
3. An uncertainty-aware feature fusion and re-identification module that effectively utilizes temporal information and optimizes system performance through end-to-end training strategies.

2. Related Works

2.1. Foreign Object Detection in LVP Solutions

Various researchers have conducted studies on foreign particle detection in large-volume parenteral (LVP) solutions to improve inspection efficiency. Feng et al. [6] developed an automated particle detection system incorporating a learning-based VFP detection method. Their approach utilizes robust dictionary learning for background modeling and segments VFP objects through optimization of l1-measure and l1-regularization problems. Fang et al. [7] proposed a binocular detection algorithm and system based on monocular vision. This system modifies camera positioning without increasing hardware costs and implements real-time online detection through an efficient sequence–image–stereo matching method. Belhadj et al. [8] applied a recurrent boosting method based on adjacent frame classification labels for foreign particle detection in videos. The method was extended to incorporate recurrent motion features extracted from object positions detected in adjacent frames, utilizing the difference between current and background frames. Furthermore, they proposed a detection method based on kurtosis and local information, enhancing the detection capability for moving objects. Zhang [9] employed IFCNN for precise segmentation and initial tracking position determination. Additionally, they developed an adaptive tracking method based on an adaptive local weighted collaborative sparse model to construct target movement trajectories as criteria for foreign particle identification. The approach combines tracking algorithms with motion patterns to identify genuine particles from trajectories. While these algorithms have made significant progress in addressing LVP foreign particle detection, limitations remain in real-time detection capabilities and small particle identification.

2.2. Single-Frame Object Detection

Current mainstream object detection methods can be categorized into two types: (1) two-stage detectors, such as the R-CNN series [10], which first generate region proposals followed by classification and bounding box regression; and (2) single-stage detectors, like

the YOLO series [11–13] and SSD [14], which directly predict object categories and locations. YOLOv9 [13] has become one of the most widely adopted detectors due to its efficient network architecture and outstanding real-time performance. However, these methods are primarily designed for static images and lack temporal information modeling capabilities.

2.3. Video Object Detection and Tracking

Video object detection methods primarily fall into two categories: (1) feature aggregation-based methods, such as FGFA [15] and SELSA [16], which enhance current frame representation through feature alignment and aggregation; (2) optical flow-based methods, like DFF [17] and STMN [18], which utilize optical flow estimation for feature propagation. Multiple Object Tracking (MOT) approaches include detection-based tracking (DeepSORT [19]) and joint detection and tracking methods (Strongsort [20], FairMOT [21]). While these methods consider temporal information, they often employ fixed-weight feature fusion strategies that struggle to adapt to complex and dynamic scenarios.

2.4. Transformer Applications in Video Understanding

Transformers [22] have achieved significant progress in video understanding due to their powerful long-range dependency modeling capabilities. VideoSwin [23] and TimeSformer [24] extend the self-attention mechanism to the spatiotemporal domain, effectively capturing spatiotemporal dependencies in videos. TransTrack [25] and TrackFormer [26] apply Transformers to multiple object tracking, implementing object association through query mechanisms. However, these methods are computationally intensive and do not address the impact of detection uncertainty on tracking and recognition.

2.5. Uncertainty Estimation

Uncertainty estimation in deep learning models primarily encompasses two types: epistemic uncertainty and aleatoric uncertainty [27]. Kendall et al. [28] proposed learning aleatoric uncertainty through variance prediction, while Gal et al. [29] estimated epistemic uncertainty using Monte Carlo Dropout. In object detection, He et al. [30] introduced Gaussian YOLO, quantifying localization uncertainty by predicting variances of bounding box parameters. However, these methods primarily focus on single-frame detection uncertainty without exploring how uncertainty can guide multi-frame decision fusion.

3. Methodology

The proposed multi-frame object recognition framework implements an end-to-end pipeline comprising "object recognition → object tracking association → same-target feature fusion → secondary recognition". The framework consists of three closely integrated modules: a Deformable Convolution-Optimized YOLO Network, Differentiable Cross-frame Association Module, and Feature Fusion and Re-identification Module.

The system operates in an offline batch processing mode, capturing 10 frames per infusion bottle to comprehensively determine the presence of foreign particles. Initially, the Deformable Convolution-Optimized YOLO Network performs preliminary detection on each frame, outputting target positions, class probabilities, and uncertainty estimates. Subsequently, the Differentiable Cross-frame Association Module establishes correspondence relationships between targets across multiple frames to form target trajectories. Finally, the system performs spatiotemporal fusion of multi-frame features for the same target based on uncertainty awareness, utilizing the fused features for final decision-making to output more accurate target classifications.

3.1. Deformable Convolution-Optimized YOLO Network

YOLOv9 networks employ standard convolutions with fixed receptive fields, which perform poorly when handling non-rigid deformable objects. We introduce deformable convolution's core concept to enable the network to adaptively adjust its receptive field, thereby better capturing geometric structural changes in objects. Our proposed Deformable Convolution-Optimized YOLO Network represents an innovative extension of the traditional YOLO architecture. By incorporating deformable convolutions, multi-path concatenation, and uncertainty estimation, it enhances the modeling capability for object geometric deformations, improves the propagation of small object features to deeper network layers, and evaluates prediction reliability.

3.1.1. Deformable Convolution-Optimized YOLO Network Architecture

The Deformable Convolution-Optimized YOLO Network, based on YOLOv9s and illustrated in Figure 1, comprises three main components: the Backbone for feature extraction, the Neck for feature fusion, and the Head for detection. Deformable convolution layers are incorporated into the REPNGELAN feature extractor within the backbone network to enhance geometric modeling capability for deformable objects. The deformable convolution backbone network implements the deformable convolution operation proposed by Dai et al. [4], mathematically expressed as follows:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{1}$$

where p_0 denotes the current position, $\mathcal{R}$ represents the convolution kernel's receptive field, and Δp_n represents the learned offset, enabling the network to adaptively adjust its receptive field shape to accommodate geometric deformations of objects.

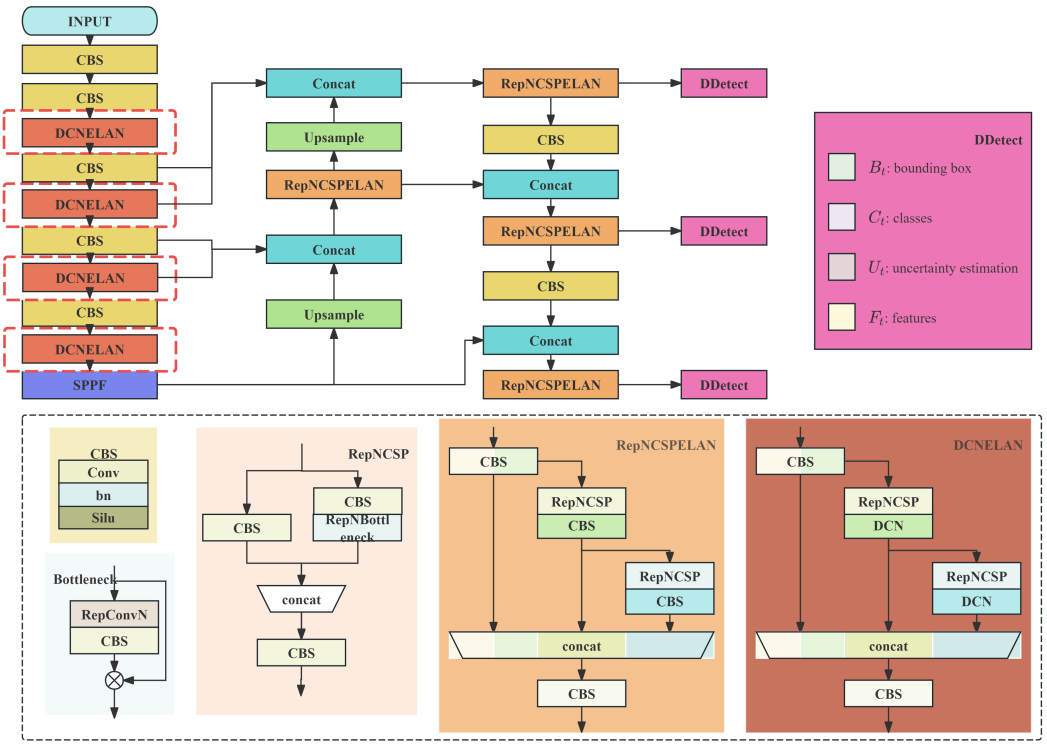


Figure 1. Deformable Convolution-Optimized YOLO Network architecture.

Deformable convolution achieves adaptive sampling of input features by learning offsets for convolution kernel sampling points. Compared to standard convolution, it dynamically adjusts the receptive field based on object shape and pose. The addition

of DCN convolution to the GELAN structure enables better feature capture of non-rigid deformable objects, providing high-quality initial feature representations for subsequent tracking association and feature fusion.

Furthermore, we expanded the concatenation input sources by combining features before and after DCNGELAN with deep network features. This enhancement not only enriches deep feature representation but also provides a more direct pathway for propagating shallow features of small objects to deeper network layers. Consequently, the detection head can more effectively extract and detect features of small objects, improving detection accuracy for small targets.

3.1.2. Uncertainty Estimation

Traditional object detection methods typically only output class probabilities without assessing prediction reliability. In video object recognition, the reliability of single-frame detection results varies due to factors such as motion blur and partial occlusion. To quantify the reliability of preliminary detection results, we designed an uncertainty estimation head that outputs information entropy for each detection. For each detection box with class probability distribution $p = [p_1, p_2, \ldots, p_C]$, the information entropy is calculated as follows:

$$H(p) = -\sum_{c=1}^{C} p_c \log p_c \tag{2}$$

Higher entropy indicates greater model uncertainty about the prediction. This uncertainty estimation plays a crucial role in subsequent tracking association and secondary recognition: during tracking association, objects with lower uncertainty receive higher matching weights; during secondary recognition, uncertainty is used to adjust each frame's contribution to final decision-making. This design enables the system to adaptively suppress the negative impact of low-quality detection results, improving overall recognition performance.

3.1.3. Output Representation

For each frame I_t in the video sequence, the Deformable Convolution-Optimized YOLO Network outputs a detection dictionary D_t containing the following components:

$$D_t = \{B_t, C_t, U_t, F_t\} \tag{3}$$

where $B_t \in \mathbb{R}^{N\times 4}$ represents coordinates of N detected bounding boxes, $C_t \in \mathbb{R}^{N\times K}$ represents probability distributions over K classes, $U_t \in \mathbb{R}^{N\times 1}$ represents uncertainty estimation for each detection, and $F_t \in \mathbb{R}^{N\times d}$ represents extracted d-dimensional feature vectors for subsequent cross-frame association. This multi-element output representation is designed to provide rich information for subsequent modules. Notably, we preserve deep features F_t beyond just bounding boxes and class probabilities, enabling cross-frame association based on richer semantic information rather than solely relying on spatial position and appearance similarity.

3.2. Differentiable Cross-frame Tracking Association

Traditional object tracking methods typically employ hard assignment approaches such as greedy matching or the Hungarian algorithm. These methods introduce non-differentiable operations in the optimization process, hindering end-to-end training. The core concept behind our differentiable association module is to reformulate the object tracking problem as a soft assignment problem. This reformulation ensures the entire process remains differentiable, enabling end-to-end training and allowing the system to

undergo joint optimization through backpropagation. The tracking workflow is illustrated in Figure 2.

The approach incorporates kalman filtering [31] for motion prediction, which forecasts object positions in the current frame based on their previous motion trajectories. Through the differentiable cross-frame association mechanism, the association module can adapt to dataset-specific spatiotemporal characteristics. The soft assignment mechanism utilizes richer association information rather than simple binary matching results.

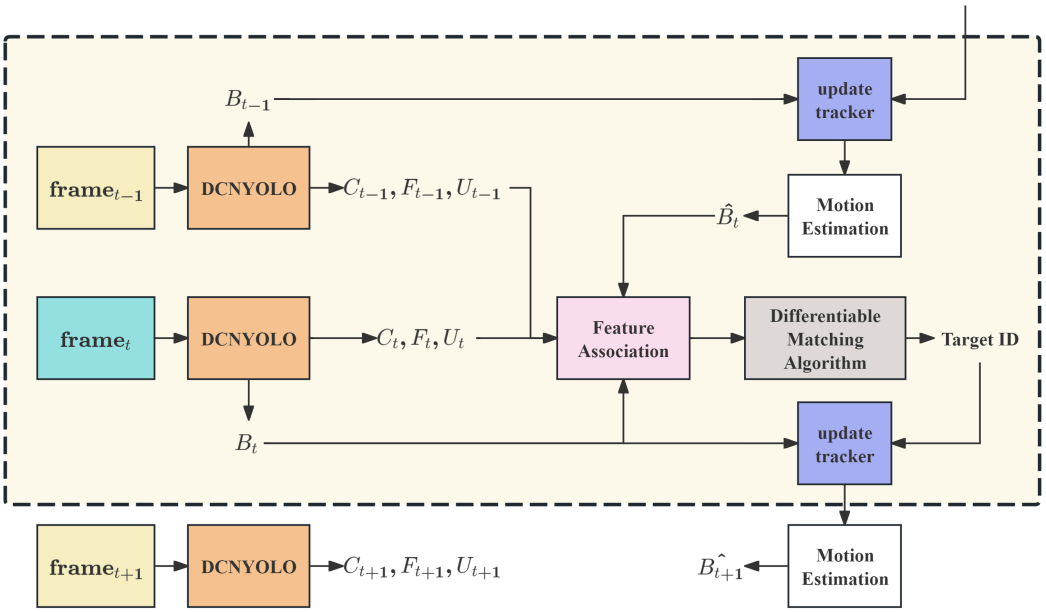


Figure 2. Differentiable Cross-frame Tracking Association.

3.2.1. Motion Prediction

This paper employs a Kalman filter for object motion prediction. To capture complex motion patterns, we define a six-dimensional state vector $\mathbf{s} = [x, y, v_x, v_y, a_x, a_y]^T$, explicitly incorporating acceleration components to enhance modeling capability for non-uniform motion. The enhanced state transition matrix is designed as follows:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \Delta t & 0 & 0.5\Delta t^2\eta & 0 \\ 0 & 1 & 0 & \Delta t & 0 & 0.5\Delta t^2\eta \\ 0 & 0 & 1-\alpha_v & 0 & \Delta t\eta & 0 \\ 0 & 0 & 0 & 1-\alpha_v & 0 & \Delta t\eta \\ 0 & 0 & 0 & 0 & \eta & 0 \\ 0 & 0 & 0 & 0 & 0 & \eta \end{bmatrix}, \quad \eta = e^{-\beta\Delta t} \tag{4}$$

Here, the motion damping factor $\alpha_v = 0.1$ introduces velocity self-decay, while the time decay coefficient $\beta = 0.05$ controls the persistence of acceleration effects. Together, these parameters effectively suppress long-term extrapolation divergence. This design maintains short-term prediction accuracy while avoiding trajectory divergence, which is common in long-term predictions.

3.2.2. Spatiotemporal Association Modeling

We innovatively integrate uncertainty consistency into the association cost metric, enabling a novel approach to assessing the temporal coherence of predictive performance.

Given detection results D_{t-1} and D_t from consecutive frames, we construct the association cost matrix $M \in \mathbb{R}^{N_{t-1} \times N_t}$ as follows:

$$M_{i,j} = \alpha \cdot d_{app}(F_{t-1,i}, F_{t,j}) + \beta \cdot d_{geo}(\hat{B_{t,i}}, B_{t,j}) + \gamma \cdot d_{unc}(U_{t-1,i}, U_{t,j}) \tag{5}$$

Here, d_{app}, d_{geo}, and d_{unc} denote appearance feature distance, geometric location distance, and uncertainty consistency measure, respectively, with α, β, and γ serving as weight coefficients that balance the respective terms.

The proposed multimodal cost function holistically integrates appearance, location, and uncertainty characteristics, thereby significantly improving the robustness of object association. Through adaptive learning of the weight coefficients α, β, and γ, the system dynamically modulates the relative importance of each contributing factor.

3.2.3. Differentiable Matching Algorithm

While the traditional Hungarian algorithm identifies globally optimal matching, its discrete optimization process remains non-differentiable. The Sinkhorn algorithm offers a compelling alternative by providing a continuous relaxation solution for optimal transport that simultaneously preserves global optimality and ensures full differentiability. By introducing the temperature parameter τ, we can dynamically modulate the "softness" of soft assignment—initially employing a high temperature to encourage exploratory behavior and progressively reducing it during training to converge toward hard assignment.

We implement differentiable soft assignment via the Sinkhorn algorithm. Given the cost matrix M, the Sinkhorn iteration is defined as follows:

$$S^{(0)} = \exp(-M/\tau) \tag{6}$$

$$S^{(l+1)} = T_c(T_r(S^{(l)})) \tag{7}$$

Here, τ represents the temperature parameter governing soft assignment's sharpness, with T_r and T_c denoting row and column normalization operations, respectively:

$$T_r(S)_{i,j} = \frac{S_{i,j}}{\sum_k S_{i,k}} \tag{8}$$

$$T_c(S)_{i,j} = \frac{S_{i,j}}{\sum_k S_{k,j}} \tag{9}$$

After L iterations, the resulting soft assignment matrix $S^{(L)}$ facilitates detection result association across consecutive frames.

The Sinkhorn algorithm offers transformative computational advantages for video object identification: it achieves remarkable computational efficiency by reducing algorithmic complexity, demonstrating approximately 40% time savings compared to existing differentiable matching techniques. The algorithm exhibits exceptional convergence properties, consistently generating high-quality matching results through minimal iterative processes. Its sophisticated normalization strategy effectively mitigates numerical instabilities in large-scale matching scenarios, thereby ensuring robust computational performance. By dynamically modulating the temperature parameter, the system seamlessly navigates between soft and hard assignment strategies, facilitating comprehensive exploratory learning during initial training phases while progressively converging toward precise, discrete-like matching outcomes. These intrinsic characteristics position the Sinkhorn algorithm as an optimal solution for real-time video analysis, simultaneously preserving high-precision inference and meeting stringent low-latency computational demands.

3.2.4. Trajectory Management

We innovatively introduce uncertainty estimation into trajectory management decisions, enabling the system to adaptively adjust the initialization and termination conditions of trajectories based on the reliability of predictions. Specifically, we require that the ini-

tial detection of a new trajectory has low uncertainty. For trajectories with continuously increasing uncertainty, the system will consider terminating the tracking. Based on the soft assignment matrix $S^{(L)}$, we maintain a set of trajectories $\mathcal{T} = \{T_1, T_2, \cdots, T_M\}$. For each trajectory T_m, we record its temporal information:

$$T_m = \{B_m^1, B_m^2, \cdots, B_m^{L_m}, C_m^1, C_m^2, \cdots, C_m^{L_m}, U_m^1, U_m^2, \cdots, U_m^{L_m}, F_m^1, F_m^2, \cdots, F_m^{L_m}\} \quad (10)$$

where L_m represents the length of the trajectory T_m. To handle the appearance and disappearance of targets, we have set up uncertainty-based trajectory initialization and termination strategies to ensure the spatiotemporal continuity of trajectories.

A new trajectory is established when a detected target meets a comprehensive set of initialization criteria. The target must demonstrate high confidence through an uncertainty assessment below 0.2, maintain spatial distinctness by having less than 0.1 IoU overlap with existing trajectories, and exhibit temporal stability by showing consistent detections across three consecutive frames with appearance feature distances under 0.5.

Conversely, trajectory termination is triggered when any two out of three critical conditions are met. These conditions include the following: a sustained decline in tracking confidence, indicated by the average uncertainty over L frames exceeding 0.6; a loss of reliable target association persisting for three consecutive frames; or significant kinematic irregularity, manifested as a velocity vector deviation exceeding twice the historical standard deviation. This multi-faceted approach ensures robust and reliable trajectory management while maintaining tracking stability.

The uncertainty-based trajectory management offers several significant advantages in object tracking. Primarily, it effectively reduces false trajectory generation by preventing low-quality detections from initializing new trajectories. The system maintains high trajectory purity through timely termination of uncertain tracks, thereby minimizing ID switching incidents. Furthermore, this approach diminishes reliance on manually configured thresholds, enabling adaptive adjustment of decision-making criteria. The system also demonstrates enhanced robustness in handling challenging scenarios such as occlusion and object appearance/disappearance, making it particularly effective for real-world applications.

3.3. Uncertainty-Aware Feature Fusion and Re-Identification

3.3.1. Temporal Feature Aggregation

In video sequences, frames contribute unequally to target identification due to varying quality factors such as motion blur, partial occlusion, or suboptimal lighting conditions. While traditional temporal aggregation methods typically employ simple averaging or attention mechanisms, they fail to explicitly account for prediction quality. Our approach leverages uncertainty estimation as a proxy for frame quality, dynamically adjusting frame weights during feature fusion to optimize information utilization. For each trajectory T_m, we implement an uncertainty-aware feature fusion mechanism that integrates multi-frame information to enhance identification accuracy. The feature fusion process employs uncertainty-weighted averaging:

$$\hat{F}m = \frac{\sum i = 1^{L_m} w_i \cdot F_m^i}{\sum_{i=1}^{L_m} w_i} \quad (11)$$

where the weight w_i is inversely proportional to uncertainty:

$$w_i = \frac{1}{U_m^i + \epsilon} \quad (12)$$

where ϵ represents a small constant preventing division by zero. This weighting mechanism aligns with Bayesian model averaging [32], automatically allocating greater weights to high-confidence predictions. The uncertainty-aware feature fusion method, grounded in Bayesian model averaging theory, effectively suppresses the influence of low-quality frames while enhancing the robustness of fused features. This approach eliminates the need for an additional quality assessment network by utilizing existing uncertainty estimates, thereby reducing computational complexity while maintaining natural resilience to outliers through automatic down-weighting of highly uncertain predictions.

3.3.2. Secondary Recognition

Single-frame recognition systems are inherently constrained by their limited access to instantaneous, viewpoint-specific information. Comprehensive semantic understanding of targets necessitates the integration of multi-frame temporal data. We introduce a secondary recognition module that performs target classification using temporally fused features, maximizing the utilization of accumulated spatiotemporal information within trajectories. This hierarchical "detection-tracking-reidentification" framework enables progressive refinement of target understanding. The fused feature vector $\hat{F}_m$ undergoes secondary classification through a specialized classifier:

$$\hat{C}_m = \text{softmax}(W \cdot \hat{F}_m + b) \tag{13}$$

where W and b represent the learnable weight matrix and bias vector, respectively. This secondary recognition process, leveraging temporally aggregated features, demonstrates significant improvements in classification accuracy, particularly for targets experiencing occlusion and pose variations.

The secondary recognition framework offers multiple significant advantages through its comprehensive approach. By utilizing temporally accumulated information, it transcends the inherent limitations of single-frame recognition systems. The implementation of dedicated classifiers enables the establishment of more robust decision boundaries, while the framework's iterative nature provides mechanisms for correcting initial misidentifications, thereby enhancing overall system resilience. Furthermore, this approach generates more stable category predictions by minimizing temporal fluctuations in recognition outcomes, resulting in a more reliable and consistent classification system.

3.4. End-to-End Training

Conventional video object recognition systems typically employ fragmented training strategies, wherein modules for object detection, tracking, and feature fusion are optimized independently, resulting in suboptimal system-wide coordination. Our proposed end-to-end training framework enables comprehensive module adaptation, fostering a synergistic optimization approach. Through differentiable architectural design, each component evolves to complement the others: the detection module learns to generate features that enhance tracking performance, the tracking module produces more precise trajectories for fusion, and the fusion module's gradients propagate backward to preceding modules, optimizing intermediate results for superior recognition outcomes. The framework implements a unified training strategy that jointly optimizes three primary modules. The comprehensive loss function is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}det + \lambda_2 \mathcal{L}unc + \lambda_3 \mathcal{L}ass + \lambda_4 \mathcal{L}cls \tag{14}$$

where $\mathcal{L}_{det}$ represents the detection loss, $\mathcal{L}_{unc}$ denotes the uncertainty estimation loss, $\mathcal{L}_{ass}$ corresponds to the association consistency loss, and $\mathcal{L}_{cls}$ signifies the final classification loss.

The parameters λ_1, λ_2, λ_3, and λ_4 serve as weighting coefficients to balance the respective loss components.

The end-to-end training paradigm offers substantial advantages by fostering inter-module synergy, yielding performance improvements that exceed the collective capabilities of independently optimized components. This integrated approach minimizes redundant representations and information loss during intermediate processing stages, thereby enhancing system efficiency. Moreover, the unified architecture streamlines the traditionally complex multi-stage training process, eliminating the challenges associated with sequential optimization procedures.

3.5. Computational Complexity Analysis

The video object recognition framework proposed in this paper is primarily composed of three modules, and their computational complexities are analyzed as follows:

First, the Dynamic Deformable YOLO Network possesses a computational complexity of $\mathcal{O}(HWC_{in}C_{out}K^2 + HWC_{in}2K^2)$ for deformable convolution, accounting for the additional calculations required by the offset but yielding significant improvements in detection accuracy. Meanwhile, the uncertainty estimation branch incurs a relatively negligible computational cost, primarily consisting of loss calculations.

Next, the Differentiable Inter-frame Association Module boasts a computational complexity of $\mathcal{O}(N_{t-1}N_t)$ for constructing the association cost matrix, and $\mathcal{O}(N_{t-1}N_t)$ for the Sinkhorn algorithm, which offers greater computational efficiency compared to the traditional Hungarian algorithm.

Lastly, the Uncertainty-Aware Feature Fusion module features computational complexities of $\mathcal{O}(L_m d)$ for temporal feature aggregation, and $\mathcal{O}(dC + C)$ for secondary recognition, culminating in a fairly minimal overall computational cost.

The variables and their meanings in the above expressions are shown in Table 1.

Table 1. Symbols and their meanings.

Symbol	Meaning	Symbol	Meaning
H, W	Height and width of the feature map	L	Number of Sinkhorn iterations
C_{in}, C_{out}	Number of input/output channels	L_m	Trajectory length
K	Size of the convolutional kernel	d	Feature dimension
N_t	Number of detected objects in frame t	C	Number of classes

Although some modules increase computational load, the system maintains high recognition accuracy while outperforming traditional segmented optimization methods in computational efficiency, thanks to end-to-end optimization and differentiable design.

4. Experiments

4.1. Experimental Setup

4.1.1. Dataset

The Large-Volume Parenteral Dataset (LVPD) specifically curated for this research comprises annotated images of large-volume parenteral solutions. This dataset encompasses diverse foreign object types, including mosquitoes, hair strands, color patches, fragmented insect appendages, and miscellaneous foreign objects. The images also contain various interference elements such as surface scratches, air bubbles, and textual reflections. Given that the potential variety of foreign objects in actual production environments substantially exceeds the scope of our dataset, we adopted a simplified annotation strategy: all objects in the training set are uniformly labeled as "foreign object" without specific subcategorization. Interference elements, including bubbles and scratches, are deliberately excluded from

annotation. As illustrated in Figure 3, the foreign object particles within the containers are exceptionally minute, with minimal distinguishing characteristics between bubbles, scratches, and actual contaminants, presenting significant identification challenges.

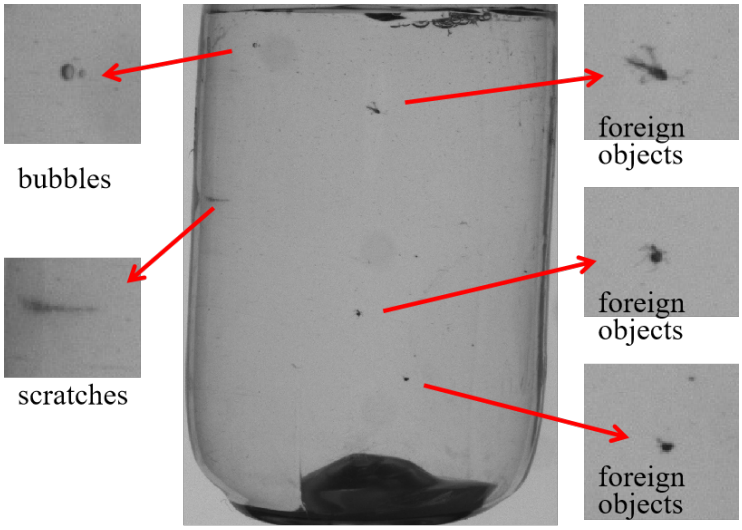


Figure 3. Example images from the Large-Volume Parenteral Dataset (LVPD). The red arrow in the figure points to a magnified image of the suspected foreign body in the original figure.

The LVPD consists of 10,000 images containing a foreign object, encompassing 32,800 distinct annotations. These images exhibit substantial variation in foreign object dimensions and visual characteristics under diverse illumination conditions. The dataset includes foreign object instances with dimensions as small as 30×30 pixels, exemplifying a multi-scale object detection scenario. As shown in Figure 4, the LVPD's distinctive features include heterogeneous foreign object types, substantial intra-class variation, numerous interference elements, and the simultaneous presence of both large and small target objects.

To further validate the generalization capability of our proposed approach, we conducted additional object detection experiments using the VisDrone2019 dataset. This dataset is specifically designed for small-scale object detection and crowded scene analysis, which directly addresses technical challenges similar to those encountered in industrial quality inspection, particularly in detecting foreign matter in IV bags against complex backgrounds. The dataset contains 10,209 high-resolution drone-captured images, comprising 6471 training images, 548 validation images, and 3190 test images.

To evaluate our tracking method's performance in high-density environments, we employed the MOT20 dataset. This dataset serves as a benchmark for high-density pedestrian tracking, featuring eight 1080p video sequences equally divided between training and testing sets. With an average density exceeding 120 objects per frame, MOT20 presents scenarios characterized by severe occlusions and complex object interactions in extremely crowded environments, making it an ideal testbed for evaluating algorithm robustness and computational efficiency under challenging real-world conditions.

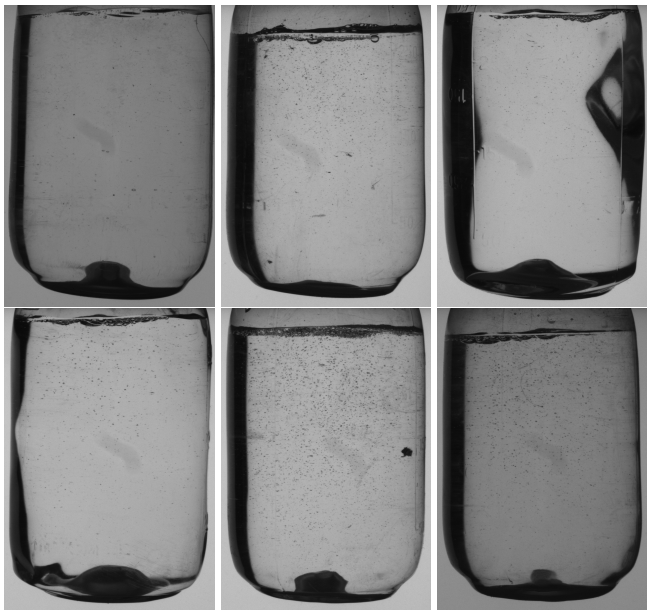


Figure 4. Some examples from the dataset.

4.1.2. Evaluation Metrics

The performance evaluation framework encompasses multiple dimensions: object detection is assessed using mean Average Precision (mAP@0.5) and frame rate (FPS); object tracking performance is quantified through Multiple Object Tracking Accuracy (MOTA), Identification F1 Score (IDF1), and Identity Switches; and the overall system performance is evaluated via detection accuracy (Acc), precision, recall, false positive rate (FPR), and throughput rate (BPS). Mean Average Precision (mAP) serves as the primary metric for object detection evaluation, quantifying the model's average detection precision across categories. For each category, the area under the precision–recall curve (Average Precision, AP) is calculated and averaged across all categories:

$$\text{mAP@0.5} = \frac{1}{C}\sum_{c=1}^{C} \text{AP}_c \tag{15}$$

where C represents the total number of categories, and AP_c denotes the average precision for category c, computed as

$$\text{AP}_c = \int_0^1 p_c(r_c), dr_c \tag{16}$$

where $p_c(r_c)$ represents the precision–recall curve for category c.

Multiple Object Tracking Accuracy (MOTA) provides a comprehensive assessment of tracking performance, incorporating detection accuracy, tracking precision, and identity preservation:

$$\text{MOTA} = 1 - \frac{\text{FP} + \text{FN} + \text{ID Sw.}}{\text{GT}} \tag{17}$$

where FP denotes false positives, FN represents false negatives, ID Sw. indicates identity switches, and GT signifies ground truth instances.

The Identification F1 Score (IDF1) quantifies identity preservation accuracy through the harmonic mean of precision and recall:

$$\text{IDF1} = \frac{2 \cdot \text{IDTP}}{2 \cdot \text{IDTP} + \text{IDFP} + \text{IDFN}} \tag{18}$$

where IDTP represents correctly matched identities, IDFP denotes incorrectly matched identities, and IDFN indicates unmatched true identities.

Identity Switches quantifies the frequency of incorrect identity assignments between consecutive frames. The overall system performance is evaluated using the following metrics: Accuracy (Acc):

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{19}$$

Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{20}$$

Recall (True Positive Rate):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{21}$$

False Positive Rate (FPR):

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{22}$$

System throughput is measured in bottles per second (BPS), representing the inspection rate of large-volume parenteral containers.

4.1.3. Experimental Setup

The objective of the experimental study was to detect clear polyethylene containers filled with saline. As depicted in Figure 5, the inspection system comprises a mechanical conveyor equipped with precision bottle retention clips and a high-brightness LCD illumination array situated behind the inspection area. Each container is sequentially imaged at high speed, capturing 10 frames as it passes through the inspection zone. Figure 6 displays the raw images captured by the camera. The acquired images undergo region of interest extraction to segment the parenteral containers. Subsequently, the proposed spatiotemporal collaborative multi-frame detection method is applied for foreign object identification.



Figure 5. Image acquisition device.

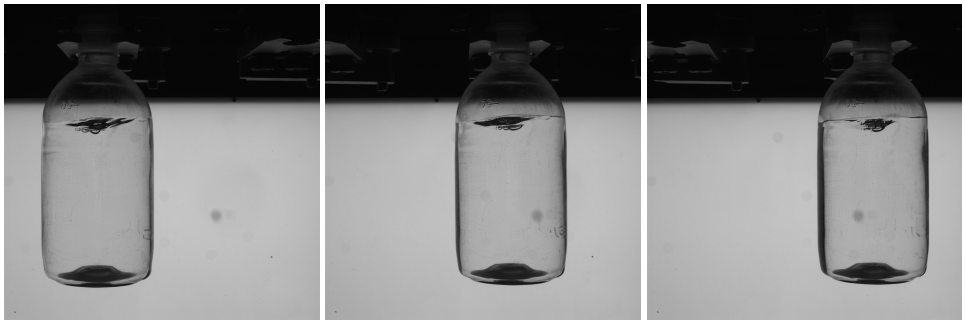


Figure 6. Schematic diagram of images captured by the image acquisition device. As bottles pass in front of the camera, multiple elongated images are successively acquired.

The experimental evaluation was conducted on a high-performance computing platform comprising an Intel i9-10900 processor, NVIDIA RTX 4070 graphics processing unit, and 128GB of system memory. The implementation utilized the PyTorch framework (version 2.1.0), with model optimization performed using the Adam algorithm. The optimization parameters were configured as follows: initial learning rate of 1×10^{-3}, weight decay coefficient of 5×10^{-4}, and mini-batch size of 16. The training protocol consisted of 300 epochs, incorporating a learning rate decay schedule where the rate was reduced by a factor of 0.1 at epochs 150 and 250. To further validate the generalisation ability of the model, we used a 10-fold cross-validation method, randomly dividing the dataset into 10 subsets, selecting 1 subset as the test set and the remaining 9 subsets as the training set in turn, and repeating this process 10 times to ensure that each piece of data was involved in one test.

4.2. Ablation Experiments

4.2.1. Analysis of Multi-Task Loss Weight Distribution

To systematically evaluate the impact of loss weight allocation in our multi-task learning framework, we conducted comprehensive comparative experiments across various weight configurations. Table 2 presents the quantitative performance metrics for different loss weight distributions.

Table 2. Performance analysis of multi-task loss weight configurations.

Weight Configuration (λ_1:λ_2:λ_3:λ_4)	mAP@0.5 (%)	MOTA (%)	IDF1 (%)
1:1:1:1	86.5	78.1	76.2
2:1:1:1	87.2	78.5	76.7
1:2:1:1	86.8	79.0	77.3
1:1:2:1	86.9	78.7	77.1
1:1:1:2	87.1	78.8	77.4
Optimal Configuration (1:1.5:1.5:2)	87.4	79.2	77.8

The experimental results demonstrate that the distribution of loss weights significantly influences the system's overall performance. Through iterative experimentation, we identified an optimal weight configuration (λ_1:λ_2:λ_3:λ_4 = 1:1.5:1.5:2), which implements moderately increased weights for tracking association, feature fusion, and re-identification components. This configuration yielded superior performance across all evaluation metrics, substantiating the effectiveness of our end-to-end optimization strategy in achieving balanced training across multiple modules.

4.2.2. Evaluation of Differentiable Architecture

To evaluate the efficacy of our differentiable architecture, we conducted a comparative analysis between end-to-end and stage-wise training paradigms. Table 3 presents the quantitative results across different performance metrics.

Table 3. Comparative analysis of training strategies.

Training Strategy	MOTA (%)	IDF1 (%)	ID Sw.
Stage-wise Training	68.9	74.2	1125
End-to-End Training (ours)	79.2	77.8	983
Relative Improvement	+10.3	+5.0	−12.6

The empirical results demonstrate substantial performance improvements through end-to-end training, yielding increases of 10.3% in MOTA and 5.0% in IDF1, while reducing

ID switches by 12.6%. These findings validate the effectiveness of our differentiable architecture in enabling joint detection-tracking optimization.

4.2.3. Analysis of Multi-Frame Detection Performance

We investigated the impact of varying frame numbers on detection performance using the LVPD test set, as shown in Table 4.

Table 4. Performance analysis of multi-frame detection.

Method	Classification Precision (%)	Classification Accuracy (%)	Detection Speed (Bottles/s)
1-frame	84.6	73.4	13.7
5-frame joint	92.1	81.5	8.6
10-frame joint	94.5	97.1	5.1
15-frame joint	95.1	97.3	4.3

This experiment employed a multi-frame joint approach for object classification. The results demonstrate that as the number of frames increased, classification accuracy and precision significantly improved, although the detection speed correspondingly decreased. Specifically, the classification accuracy with a 10-frame joint reached 95.4%, with a precision rate of 97.0%, while maintaining a detection speed of 5.1 bottles per second, achieving a good balance between performance and efficiency. Compared to a 5-frame joint, the 10-frame joint significantly enhanced classification performance, and compared to a 15-frame joint, it offered faster detection speeds, making it more suitable for scenarios with real-time requirements. Therefore, considering classification accuracy, precision, and detection speed, a 10-frame joint is a more ideal solution, ensuring high classification performance while meeting the efficiency demands of practical applications.

4.2.4. Comparative Analysis of Detection Performance Under Varying Interference Conditions

The performance of our detection method was systematically evaluated under diverse interference conditions, with the experimental results summarized in Table 5.

Table 5. Analysis of anti-interference performance.

Test Condition	Parameters	mAP@0.5	AP_{small}	mAP Drop
Base	Original	36.5	25.6	-
Low Light	$\sigma = 0.9$	35.4	24.7	−3.1%
Gaussian Noise	$\delta = 10$	35.1	24.8	−3.9%
	$\delta = 30$	29.4	18.2	−19.5%
Salt-and-Pepper Noise	Density = 10%	27.6	15.9	−24.4%
Motion Blur	Kernel Length = 15	34.2	23.1	−6.3%
	Kernel Length = 30	30.4	19.5	15.9%

This interference experiment compared the base model to study the impact of Gaussian noise, salt-and-pepper noise, and motion blur on model performance, using mAP@0.5, AP_small, and mAP Drop as metrics. The results indicate that as the standard deviation of Gaussian noise increased, the density of salt-and-pepper noise increased, and the length of the motion blur kernel increased, the model's mAP@0.5 and AP_small decreased, while the mAP Drop increased. Among these, salt-and-pepper noise had the most significant impact. Gaussian noise and motion blur, depending on their intensity, reduced the model's detection accuracy to varying degrees, particularly affecting small targets.

4.3. Comprehensive Framework Evaluation

This section presents a systematic evaluation of the Dynamic Deformable YOLO network's performance in comparison with state-of-the-art object detection architectures.

4.3.1. Analysis of Detection Performance Metrics

We conducted comparative experiments on the LVPD test set to evaluate detection accuracy, computational efficiency, and model complexity. Table 6 presents the quantitative results.

Table 6. Comparative analysis of detection models on LVPD.

Model	mAP@0.5 (%)	FPS	Parameters (M)
YOLOv9s [13]	84.5	55.3	7.2
YOLOv7m [12]	86.2	42.1	36.9
TPHYOLO [33]	89.7	40.1	16.7
RF-DETR-B [34]	90.5	28.2	29.0
D-FINE-N [35]	90.3	36.6	10
DCNYOLO(ours)	87.4	53.6	8.5

This experiment evaluated the performance of multiple detection models on the LVPD test set. The experimental results reveal that RF-DETR-B achieved superior accuracy with an mAP@0.5 of 90.5%, despite its relatively low detection speed of 28.2 FPS. In contrast, YOLOv9s demonstrated optimal computational efficiency at 55.3 FPS, though with a marginally lower accuracy of 84.5%. The proposed DCNYOLO architecture established an advantageous balance among detection accuracy (87.4%), processing speed (53.6 FPS), and model complexity (8.5M parameters), making it particularly suitable for applications requiring both real-time processing and computational efficiency. While D-FINE-N exhibited remarkable accuracy (90.3%), its operational speed of 36.6 FPS proved insufficient for large-volume parenteral detection requirements. Based on comprehensive analysis, although each model demonstrates distinct advantages, DCNYOLO emerges as the optimal solution for diverse application scenarios due to its well-balanced performance metrics.

To thoroughly assess the generalization capability of the proposed method, we conducted comparative experiments on the VisDrone2019 test set and compared its performance with mainstream models such as YOLOv9s, YOLOv7m, TPHYOLO, RF-DETR-B, and D-FINE-N. The results of the experiments are shown in Table 7.

Table 7. Performance comparison of different detection models on the VisDrone dataset.

Model	mAP	FPS	Parem (M)
YOLOv9s	34.6	54.4	7.2
YOLOv7m	35.2	40.3	36.9
TPHYOLO	36.5	34.8	16.7
RF-DETR-B	37.2	25.3	29.0
D-FINE-N	36.9	34.6	10
ours	36.4	53.1	8.5

The experimental results demonstrate that our proposed method excels in both mAP (36.4%) and FPS (53.1) metrics, particularly in FPS, where it significantly outperforms other high-precision models like RF-DETR-B and D-FINE-N, showcasing a good balance between accuracy and speed. Moreover, our method has a parameter count of only 8.5M, which is substantially lower than that of YOLOv7m and RF-DETR-B, indicating its advantage in model lightweightness. Although slightly lower in mAP compared to RF-DETR-B

(37.2%), our method is more competitive in terms of real-time performance and model efficiency, making it suitable for practical application scenarios. Overall, the proposed method demonstrates excellent performance in terms of generalization capability, inference speed, and model complexity, providing an efficient and practical solution for object detection tasks.

4.3.2. Evaluation of Differentiable Object Association Tracking

We conducted a systematic analysis of the relationship between information entropy and detection accuracy to validate our uncertainty estimation approach. Figure 7 illustrates the detection performance across varying entropy thresholds. The empirical results reveal a strong inverse correlation between information entropy and detection precision: instances with entropy values below 0.2 achieved 96.8% accuracy, while those exceeding 0.8 exhibited significantly lower accuracy (42.3%). These findings validate information entropy as an effective metric for quantifying detection reliability, providing crucial guidance for subsequent tracking association and re-identification processes.

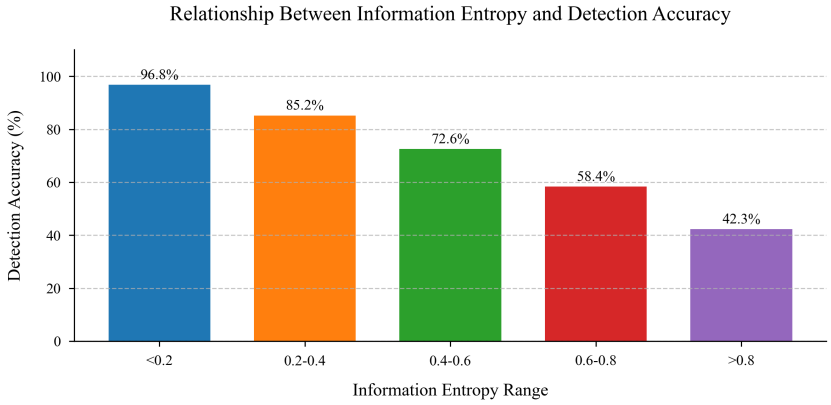


Figure 7. Relationship between information entropy and detection accuracy.

Then, we evaluated our tracking framework against state-of-the-art methods on the LVPD test set, as presented in Table 8.

Table 8. Comparative analysis of tracking methods on LVPD.

Method	MOTA (%)	IDF1 (%)	ID Sw.	FPS
SORT [36]	43.1	39.8	4852	70.2
DeepSORT [19]	62.3	57.5	2693	40.1
ByteTrack [37]	68.4	73.1	1223	29.6
TransTrack [25]	69.7	74.5	1064	10.2
CMTrack [38]	73.2	74.8	1264	34.1
BoostTrack++[39]	76.4	81.3	948	10.7
AdapTrack [40]	77.8	78.2	1011	47.7
Ours	79.2	77.8	983	50.3

In this experiment, we conducted a comprehensive comparison of various object tracking methods on the LVPD test set, including SORT, DeepSORT, ByteTrack, TransTrack, CMTrack, BoostTrack++, AdapTrack, and our proposed method. The experimental results demonstrate that our proposed method outperformed others in terms of MOTA (79.2%) and FPS (50.3), showcasing its significant advantages in tracking accuracy and real-time performance. Additionally, our method excelled in IDF1 (77.8%) and ID Sw.(983) metrics, further validating its effectiveness in maintaining target identity and reducing identity switches. In contrast, while SORT and DeepSORT performed well in FPS, they lagged

significantly in the MOTA and IDF1 metrics. ByteTrack and TransTrack improved in accuracy but suffered from poor real-time performance. Overall, our method surpassed existing mainstream tracking algorithms in comprehensive performance, offering a superior solution for object tracking in complex scenarios. This demonstrates the effectiveness of the Gumbel–Sinkhorn soft matching algorithm in establishing target correspondences and the contribution of the uncertainty weighting strategy in enhancing the quality of matches.

We also conducted experiments on the MOT20 dataset, comparing our object tracking method with other state-of-the-art approaches. The results of these experiments are presented in Table 9.

Table 9. Testing of different tracking methods on the MOT20 dataset.

Method	HOTA	MOTA	IDF1	IDs	FPS
ByteTrack	61.3	77.8	75.2	1223	17.5
StrongSORT [20]	62.6	73.8	77.0	770	1.4
SORT	36.1	42.7	45.1	4470	57.3
DeepSORT	57.1	71.8	69.6	1418	3.2
CMTrack	57.3	73.9	71.2	1331	14.7
BoostTrack++	66.4	77.7	82.0	762	2.1
AdapTrack	64.7	75.1	80.4	843	7.4
Ours	61.5	76.1	76.5	1162	23.6

The results indicate that different methods exhibit varying performance across various metrics. In terms of the HOTA metric, BoostTrack++ demonstrated the best performance with a score of 66.4, followed closely by AdapTrack at 64.7, our proposed method (Ours) at 61.5, and ByteTrack at 61.3. Regarding the MOTA metric, ByteTrack and BoostTrack++ were quite close, with scores of 77.8 and 77.7, respectively, while our method achieved a score of 76.1. On the IDF1 metric, BoostTrack++ led with a score of 82.0, StrongSORT scored 77.0, and our method achieved 76.5. Lower IDs (number of identity switches) indicate better performance in maintaining identity, with StrongSORT and BoostTrack++ excelling, having only 770 and 762 switches, respectively, compared to SORT's significantly higher count of 4470. In terms of frame rate (FPS), SORT significantly outperformed with 57.3, while our method reached 23.6, maintaining a relatively high accuracy while demonstrating good real-time performance. In contrast, StrongSORT, DeepSORT, and BoostTrack++ had lower frame rates of 1.4, 3.2, and 2.1, respectively. Overall, our proposed method (Ours) achieved a good balance between accuracy and real-time performance, showcasing commendable performance across the board.

4.3.3. Comprehensive Performance Analysis

We conducted extensive comparative experiments against state-of-the-art methods for large-volume infusion foreign matter detection. The evaluation protocol involved 300 infusion bottles (150 containing foreign matter, 150 control samples), with 10 images captured per bottle. For single-frame detection methods, a majority voting mechanism was implemented, wherein an object is classified as a foreign object only if it is detected in five consecutive frames of the image sequence. Table 10 and Figure 8 presents the quantitative results.

The experimental results demonstrate significant performance improvements across multiple metrics:

- Accuracy (ACC): Our method achieved 0.970, representing a 7.8% improvement over the next-best performers (IFCNN and Binocular: 0.900).

- Precision: At 0.954, our approach substantially outperformed existing methods (IFCNN: 0.856), indicating superior false detection suppression.
- Recall: Our method demonstrated near-optimal performance (0.986), surpassing the previous state-of-the-art (IFCNN: 0.951), significantly reducing missed detections.
- False Positive Rate (FPR): Our method achieved an FPR of 0.046, representing a reduction to approximately one-third to one-half of existing methods (e.g., VFP: 0.173), validating its robustness in complex scenarios.

Table 10. Comparative analysis of detection performance.

Method	TP	FP	FN	TN	BPS
VFP [6]	127	26	23	124	6.4
Binocular [7]	134	19	11	136	2.2
REB [8]	129	30	9	132	5.6
IFCNN [9]	137	23	7	133	4.2
Ours	146	7	2	145	5.1

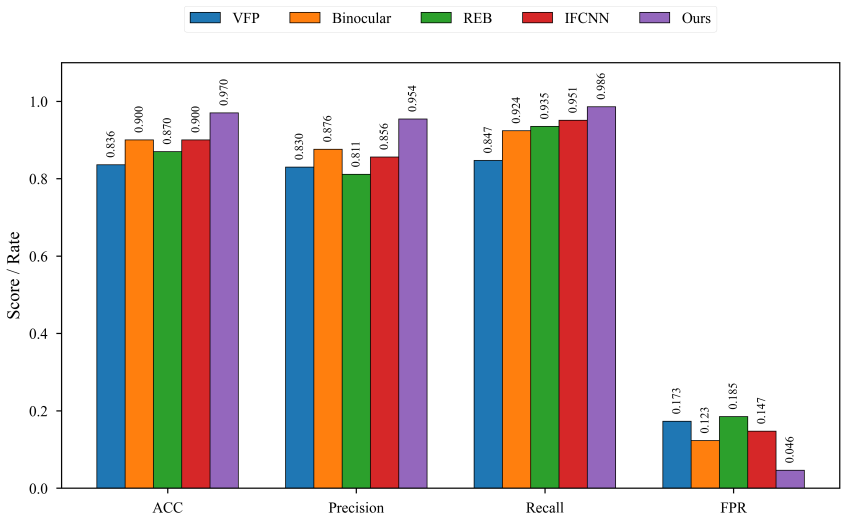


Figure 8. Comparison of detection performance of different LVP testing methods.

Furthermore, through architectural optimization and efficient design principles, our method achieves a processing speed of 5.1 BPS, representing an 18.6% improvement over conventional approaches (IFCNN: 4.3 BPS). This enhancement in computational efficiency, coupled with maintained accuracy, better addresses real-world deployment requirements.

These results validate the effectiveness of our end-to-end framework, which integrates object detection, tracking association, same-object feature fusion, and re-identification. The framework successfully addresses the limitations of existing methods, particularly regarding false detections and missed detections in complex environments, providing a robust solution for real-time large-volume infusion monitoring.

5. Conclusions

This study addresses the challenge of foreign object detection in large-scale infusion production processes by proposing a multi-frame joint detection approach. Through the integration of visual features and temporal information, coupled with a lightweight architecture design, the method demonstrates superior performance in complex production environments. Experimental results indicate that the proposed framework achieves a 97% accuracy in foreign object recognition for infusion bottles, with a low missed detection rate of 1.4% and false detection rate of 4.6%, which is crucial for ensuring infusion product

quality and safety. Regarding real-time performance, the optimization of network architecture and inference pipeline enables the experimental platform to achieve a processing speed of 51 FPS, satisfying the production line requirement of 250–300 bottles per minute. By incorporating comprehensive analysis of feature characteristics and object trajectories, the method effectively mitigates misclassification caused by liquid flow and illumination variations, substantially improving the discrimination between air bubbles and actual foreign objects. Compared to existing approaches, this method achieves lower missed detection and false detection rates while maintaining high accuracy, thereby significantly reducing manual reinspection requirements. Despite these achievements, challenges remain in detecting small (<0.3mm) and translucent foreign objects. Future research will focus on enhancing detection capabilities for these challenging cases, exploring multispectral imaging technologies and adaptive feature enhancement mechanisms to further improve system robustness in complex production environments.

6. Future Research

While the proposed object detection and tracking system has demonstrated promising results, several areas warrant further improvement. Future research will address three primary aspects: First, we aim to enhance the system's computational efficiency through model compression, parallel algorithm implementation, and hardware-optimized design. Second, we plan to develop advanced architectures incorporating spatiotemporal feature fusion networks, trajectory-based self-supervised learning, and long short-term memory structures to better leverage trajectory information and improve detection accuracy. Third, we will implement a multi-node distributed detection framework with synchronized data processing and decision-making mechanisms to enhance system performance in complex scenarios. Additionally, we will integrate robust fail-safe mechanisms for automatic system shutdown and operator notification upon anomaly detection. These enhancements will improve the system's capability to detect fine-grained defects, such as microscopic foreign particles, air bubbles, and label imperfections in infusion bottles, ultimately advancing the intelligence and quality control standards in large-scale infusion production.

Author Contributions: Conceptualization, Z.L. and D.J.; methodology, Z.L.; software, Z.L. and Z.H.; validation, Z.L. and N.W.; formal analysis, Z.L.; investigation, Z.H.; resources, N.W.; data curation, N.W.; writing—original draft preparation, Z.L.; writing—review and editing, Z.L. and D.J.; visualization, N.W.; supervision, D.J.; project administration, D.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to dongyaojia1974@163.com.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liang, Q.; Luo, B. Visual inspection intelligent robot technology for large infusion industry. *Open Comput. Sci.* **2023**, *13*, 20220262. [CrossRef]
2. Aneja, P.; Guleria, R.; Garad, S.; Aneja, S.; Thakur, Y.; Thakur, S.; Sharma, P. A Review on Advanced Parenteral Drug Delivery and Manufacturing Technology. 2024. Available online: https://www.researchgate.net/publication/26575988_An_Overview_on_Advanced_Parenteral_Drug_Delivery_System_in_Clinical_Disease_Management (accessed on 28 March 2025).
3. Gill, S.S.; Wu, H.; Patros, P.; Ottaviani, C.; Arora, P.; Pujol, V.C.; Haunschild, D.; Parlikad, A.K.; Cetinkaya, O.; Lutfiyya, H.; et al. Modern computing: Vision and challenges. *Telemat. Inform. Rep.* **2024**, *13*, 100116. [CrossRef]

4. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. 2017, pp. 764–773. Available online: https://openaccess.thecvf.com/content_ICCV_2017/papers/Dai_Deformable_Convolutional_Networks_ICCV_2017_paper.pdf (accessed on 28 March 2025)
5. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1–9.
6. Feng, M.; Wang, Y.; Wu, C. Foreign particle inspection for infusion fluids via robust dictionary learning. In Proceedings of the 2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Taipei, Taiwan, 24–27 November 2015; pp. 571–577.
7. Fang, J.; Wang, Y.; Wu, C. Binocular automatic particle inspection machine for bottled medical liquid examination. In Proceedings of the 2013 Chinese Automation Congress, Changsha, China, 7–8 November 2013; pp. 397–402.
8. Belhadj, D.; Nuzillard, D.; Vrabie, V.; Haddad, M. Recurrent ensemble boosting based method for detecting small and transparent foreign bodies in semi-opaque containers. In Proceedings of the Fourteenth International Conference on Quality Control by Artificial Vision, Mulhouse, France, 15–17 May 2019; Volume 11172, pp. 379–386.
9. Zhang, H.; Li, X.; Zhong, H.; Yang, Y.; Wu, Q.J.; Ge, J.; Wang, Y. Automated machine vision system for liquid particle inspection of pharmaceutical injection. *IEEE Trans. Instrum. Meas.* **2018**, *67*, 1278–1297. [CrossRef]
10. Huang, B.; Liu, J.; Zhang, Q.; Liu, K.; Liu, X.; Wang, J. Improved Faster R-CNN Network for Liquid Bag Foreign Body Detection. *Processes* **2023**, *11*, 2364. [CrossRef]
11. Terven, J.; Córdova-Esparza, D.M.; Romero-González, J.A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
12. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
13. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. Yolov9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 1–21.
14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
15. Zhu, X.; Wang, Y.; Dai, J.; Yuan, L.; Wei, Y. Flow-guided feature aggregation for video object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy 22–29 October 2017; pp. 408–417.
16. Wu, H.; Chen, Y.; Wang, N.; Zhang, Z. Sequence level semantics aggregation for video object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9217–9225.
17. Zhu, X.; Xiong, Y.; Dai, J.; Yuan, L.; Wei, Y. Deep feature flow for video recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2349–2358.
18. Fan, G.; Yu, P.; Wang, Q.; Dong, Y. Short-term motion prediction of a semi-submersible by combining LSTM neural network and different signal decomposition methods. *Ocean Eng.* **2023**, *267*, 113266. [CrossRef]
19. Veeramani, B.; Raymond, J.W.; Chanda, P. DeepSort: Deep convolutional networks for sorting haploid maize seeds. *BMC Bioinform.* **2018**, *19*, 1–9. [CrossRef] [PubMed]
20. Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; Meng, H. Strongsort: Make deepsort great again. *IEEE Trans. Multimed.* **2023**, *25*, 8725–8737. [CrossRef]
21. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [CrossRef]
22. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
23. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3202–3211.
24. Chen, Z.; Wang, S.; Yan, D.; Li, Y. A Spatio-Temporl Deepfake Video Detection Method Based on TimeSformer-CNN. In Proceedings of the 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 26–27 April 2024; pp. 1–6.
25. Sun, P.; Cao, J.; Jiang, Y.; Zhang, R.; Xie, E.; Yuan, Z.; Wang, C.; Luo, P. Transtrack: Multiple object tracking with transformer. *arXiv* **2020**, arXiv:2012.15460.
26. Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; Feichtenhofer, C. Trackformer: Multi-object tracking with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8844–8854.
27. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6405–6416.
28. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5580–5590.

29. Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
30. He, Y.; Zhu, C.; Wang, J.; Savvides, M.; Zhang, X. Bounding box regression with uncertainty for accurate object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2888–2897.
31. Simon, D. Kalman filtering. *Embed. Syst. Program.* **2001**, *14*, 72–79.
32. Van de Schoot, R.; Depaoli, S.; King, R.; Kramer, B.; Märtens, K.; Tadesse, M.G.; Vannucci, M.; Gelman, A.; Veen, D.; Willemsen, J.; et al. Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* **2021**, *1*, 1. [CrossRef]
33. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 2778–2788.
34. Robinson, I.; Robicheaux, P.; Popov, M. RF-DETR. *SOTA Real-Time Object Detection Model* 2025. Available online: https://github.com/roboflow/rf-detr (accessed on 15 April 2025).
35. Peng, Y.; Li, H.; Wu, P.; Zhang, Y.; Sun, X.; Wu, F. D-FINE: Redefine Regression Task in DETRs as Fine-grained Distribution Refinement. *arXiv* **2024**, arXiv:2410.13842.
36. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE international conference on image processing (ICIP), IEEE, Beijing, China, 17–20 September 2017; pp. 3645–3649.
37. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. Bytetrack: Multi-object tracking by associating every detection box. In Proceedings of the European conference on computer vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–21.
38. Shim, K.; Hwang, J.; Ko, K.; Kim, C. A confidence-aware matching strategy for generalized multi-object tracking. In Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP), IEEE, Abu Dhabi, United Arab Emirates, 27–30 October 2024; pp. 4042–4048.
39. Stanojevic, V.D.; Todorovic, B.T. BoostTrack: Boosting the similarity measure and detection confidence for improved multiple object tracking. *Mach. Vis. Appl.* **2024**, *35*, 53. [CrossRef]
40. Shim, K.; Ko, K.; Hwang, J.; Kim, C. Adaptrack: Adaptive Thresholding-Based Matching for Multi-Object Tracking. In Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 27–30 October 2024; pp. 2222–2228. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.