Robust Hybrid Deep Learning and Filtering Model for Occlusion Underwater Object Tracking

Venkata Krishna Chaitanya Putrevu [1,2,*] and Chandra Bhushana Rao Kota [3]

[1] Department of Electronics and Communication Engineering,
Jawaharlal Nehru Technological University Kakinada, Kakinada, India
[2] Department of Electronics and Communication Engineering,
Gayatri Vidya Parishad College of Engineering for Women, Visakhapatnam, India
[3] Department of Electronics and Communication Engineering, UCEV,
Jawaharlal Nehru Technological University-Gurajada Vizianagaram, Vizianagaram, India
Email: chaitanyaputrevu@gmail.com (V.K.C.P.); cbraokota.ece@jntugvcev.edu.in (C.B.R.K.)
*Corresponding author

***Abstract*—Underwater object detection plays a vital role in marine exploration, surveillance, and ecological monitoring. However, the presence of light scattering, color distortion, and low visibility makes accurate detection and tracking in underwater environments extremely challenging. To address these issues, this work proposes an efficient underwater object detection and tracking framework integrating the Single Shot MultiBox Detector (SSD), Deep Simple Online and Realtime Tracking (SORT), and Kalman Filter (KF). The SSD model is employed for real-time object detection due to its capability of multi-scale feature extraction and fast inference. Detected objects are then tracked across frames using the Deep SORT algorithm, which combines appearance descriptors and motion information for robust identity preservation. To further enhance trajectory estimation and reduce noise, the Kalman Filter is incorporated to predict object positions in occluded or visually degraded frames. The proposed hybrid approach demonstrates improved stability and accuracy in underwater video sequences utilizing the benchmark Underwater Object Tracking (UOT32) dataset. Performance evaluation based on precision, recall, and F1-Score confirms the model's robustness in identifying and tracking objects under varying underwater conditions. The obtained results with a precision of 97.39%, a recall 100%, and F1-Score of 0.986 indicate that the integration of SSD with Deep SORT and Kalman filtering offers a reliable and computationally efficient solution for real-time underwater detection and tracking applications.**

***Keywords*—underwater object detection, single shot multi-box detector, deep Simple Online and Realtime Tracking (SORT), Kalman Filter (KF), object tracking**

I. Introduction

In the last ten years, many innovative approaches have been developed for underwater object identification, which has resulted in outstanding progress. These approaches' groundbreaking successes in underwater object identification are what keep the field moving forward. The development of underwater object identification tasks is being propelled forward by these approaches, which have made substantial contributions. Moreover, these accomplishments offer crucial technical assistance for a wide range of disciplines, including scientific inquiry, the development of marine resources, and many more. The tracking of objects is one of the most basic and ongoing problems in computer vision. The main objective is to create a reliable appearance model using the starting point of the target in the first frame. This model will be utilized for tracking the target in consecutive frames.

Some of the existing model designed for the detection of small underwater objects using deep learning models. Chen and Er [1] proposed a dynamic You-Only-Look-Once (YOLO) model which is a backbone with deformable convolutions and a feature fusion scheme that is channel-wise and scale-wise spatial aware to better detect small objects in underwater images.

Marine environmental engineering relies on precise underwater item detection. In complicated underwater settings, many of the suggested underwater object identification algorithms with relatively high accuracy fail to produce satisfying results due to the enormous number of parameters and Floating-Point Operations (FLOPs) involved. Enhanced feature-based underwater object identification is a one-stage approach that was suggested by Zu *et al*. [2]. The lowest FLOPs and parameters are achieved at the same time.

While there have been studies showing that underwater image enhancement methods can improve detector detection accuracy, no research has looked at how these two tasks are related. For the most part, this is since neither high-quality reference pictures nor bounding box annotations are included in the underwater datasets that are used to determine detection accuracy or image quality evaluation measures. Chen *et al*. [3] offers the WaterPairs dataset, a large-scale dataset for underwater item detection

Manuscript received November 30, 2025; revised January 16, 2026; accepted March 5, 2026; published May 29, 2026.

along with bounding box annotations and high-quality reference photos, so that we may study the effects of underwater image enhancement approaches on these tasks. Researchers may use the WaterPairs dataset to examine the effects of methods that improve underwater images on underwater item recognition tasks in a systematic way. In this data, only static images are available. To further enhance the detection of underwater objects, video sequences datasets are considered by the authors.

In underwater settings, visual distortions, color cast problems, and poor visibility circumstances add to the difficulties of the Visual Object Tracking (VOT) job, which are already problematic in open-air contexts. In response to these issues, Li *et al*. [4] present a new framework for underwater target tracking that uses a Correlation Filter (CF) in conjunction with picture improvement and a two-stage feature compression method. The model struggles to provide high-quality detection while keeping the target trajectory stable. To overcome all the issues, a novel model is designed in this work. Although well-known elements like Single Shot Detectors (SSD), Deep Simple Online and Realtime Tracking (SORT), and Kalman filtering are used, their preparation, parameterization, and integration are specifically designed for underwater video situations, where motion uncertainty and appearance deterioration differ significantly from terrestrial settings. One of the main issues in underwater tracking applications is trajectory stability and recovery under occlusion, which is the emphasis of the suggested approach.

In this paper an SSD is utilized, which provides a good trade-off between detection speed and localization accuracy. Their multi-scale nature allows detection across different object sizes, and they are often easier to deploy in resource constrained scenarios. Deep SORT is a simple online and real-time tracking with a deep association metric adds tracking, using appearance features plus motion, which helps identity consistency over time. The Kalman Filter adds predictive power to deal with occlusions, missed detections, or noisy detections, and smoothens the trajectories. However, limited work has combined all three models, i.e., SSD + Deep SORT + Kalman Filter, specifically tuned for the underwater domain, especially under the kinds of image degradations such as noise, low contrast, and color shift that characterize real underwater footage.

The main contribution of the work in the detection and tracking of underwater objects is as follows:

- Design of SSD for object detection adjusted for underwater picture settings and employs both appearance traits and motion (via bounding box changes) to track identities across frames using Deep SORT.
- Design of Kalman Filter to smooth out trajectories and predict them, which is great for situations when detections are spotty or blocked.
- The performance of the proposed model is evaluated using F1-Score, recall, and precision metrics.

This paper's remaining sections are structured as follows: Section II lists the relevant work by various researchers. Section III goes into the designing of SSD, deep sort and Kalman filter. Section IV presents experimental results, and discusses benchmark datasets. In Section V, an ablation study for the proposed model is given, and finally, Section VI ends up with a conclusion of the findings.

II. Related Work

A brief synopsis of current methods for underwater object tracking is provided in this section. Problems specific to underwater tracking are the focus of research. This article delves into the history of tracking algorithms in both surface and submerged environments by discussing various methods that have been developed to address certain problems.

Deep learning approaches are now the center of attention in underwater object identification research, which aims to improve the accuracy and universality of existing algorithms. A watershed points in deep learning's meteoric rise to prominence in object identification and detection occurred with the development of the Region-Based Convolutional Neural Network (R-CNN) [5]. Currently, a growing number of researchers are utilizing deep learning to identify objects underwater, which has led to groundbreaking discoveries. The classification of object detection techniques is currently available as either two-stage or single-stage. An example of a two-stage method is the R-CNN series, which includes Fast R-CNN [6] and Faster R-CNN [7]. These algorithms first generate region suggestions and then use these proposals to perform classification and regression tasks. These algorithms have shown better detection performance; however, they are not very efficient processors. The goal of single-stage algorithms like the YOLO series of algorithms [8] and the single-shot multi-Box detector [9] is to achieve fast detection with good performance.

To successfully detect targets at various depths and decrease radiated noise, Yang *et al.* [10] integrated a deep autoencoder neural network with a deep long short-term memory network (DLSTM). To identify and categorize ship-radiated noise, they employed a pretrained DLSTM model in conjunction with a SoftMax classifier. Changes in the status of the target or object, as well as occlusion, have a major effect on target detection. To recognize underwater pictures with blur, occlusion, and overlap, Lin *et al.* [11] presented an approach named RoIMix, which demonstrated better generalization performance. To accomplish underwater object identification, Lau and Lai [12] concentrated on refining Faster R-CNN's fundamental network architecture through selection and improvement.

To improve contextual relevance and multiscale detection capabilities, they created a composite linked backbone network that leverages the advantages of several backbone networks [13, 14]. In a single run, these algorithms can estimate the category and location of targets using direct regression approaches. To solve the problem of high ammonia nitrogen levels in aquaculture, which is caused by feed particles not being consumed,

Hu *et al*. [15] changed the network connections and substituted the feature mapping that is responsible for big features in YOLO-v4 with feature maps that are more finely grained. By doing away with unnecessary steps, their method greatly enhanced the accuracy of detection and identification in actual breeding settings.

Yuan *et al*. [16] have also suggested an additional underwater object identification technique called YOLOv5. They strengthened the algorithm's multiscale feature fusion approach and confidence loss function and used the twin transformer as the algorithm's backbone network to make it work better in underwater settings. Zhang *et al.* [17] created a compact system for underwater object identification using attention feature fusion, MobileNet v2, and YOLO-v4 algorithms. By reducing the number of factors, their suggested strategy creates a lighter model that greatly enhances detection speed and accuracy.

The image enhancement and deep network feature extraction were discussed by Chen *et al*. [18] to combat underwater picture deterioration. There is an opportunity to enhance the tracking frame's placement accuracy; nevertheless, as testing was restricted to a tiny dataset. Wang *et al.* [19] suggested a new backbone network architecture, NewNet-62. It included an inverted residual bottleneck block and the SiamRPN++ algorithm, which resulted in performance and accuracy improvements [20]. The optimisation of SiamPRN also made it a better fit for use on submerged platforms [21]. Using motion data to improve tracking, SreekalaK *et al*. [22] integrated a Kalman Filter (KF) with a deep convolutional neural network to solve the object tracking problems that come with underwater communication.

The ability to reliably associate detections across frames into persistent tracks is crucial for robust multi-object tracking. A traditional real-time tracking pipeline integrates online association and motion prediction with a per-frame detector. Deep SORT included a learnt appearance embedding to significantly decrease identity shifts by integrating motion and visual resemblance during association, whereas SORT introduced a pragmatic Kalman-filter + Hungarian assignment architecture that prioritizes simplicity and speed. Underwater trackers, which rely on these frameworks for computing efficiency, are among the numerous tracking systems that use them [23].

When images taken underwater are of low quality, underwater target identification systems struggle to do their jobs. The current state of deep learning networks is inadequate for efficient and accurate detection procedures in underwater devices because of their limited computational capability. The accuracy of target detection was improved by reducing the number of model parameters using a lightweight method that relies on multi-scale feature fusion [24]. Recent UOT research investigates domain-adapted Re-Identification (Re-ID) networks, data augmentation, and huge pretraining corpora to construct more resilient appearance models, as learnt appearance descriptors encounter domain shift difficulties in underwater situations, such as color washout and blur, which can decrease descriptor discriminability. In this regard, the publishing of extensive undersea collections is a crucial step: When direct re-identification is needed after extended occlusions, large, diverse corpora can be used to pretrain appearance embeddings that are more able to generalize across underwater circumstances and object categories, enhancing Deep-SORT-style association [25].

Existing detector improvements often stop at detection benchmarks, and many tracking studies either do not focus on occlusion-specific metrics or do not systematically ablate the role of Kalman filtering and re-association logic under underwater occlusion patterns. These gaps motivate our occlusion-aware SSD-Deep SORT-KF design and the focused empirical evaluation on Underwater Object Tracking (UOT32), where we measure both detection and tracking under representative underwater occlusion scenarios.

III. Methodology

The proposed methodology integrates Single Shot MultiBox Detector (SSD) for object detection, DeepSORT for appearance-based multi-object association, and a Kalman Filter (KF) for predictive motion modelling and occlusion recovery. This hybrid deep learning framework is designed to enhance tracking robustness in the visually challenging underwater domain, where factors such as light scattering, turbidity, and partial occlusion severely affect object visibility. Fig. 1 illustrates the overall workflow of the proposed tracking pipeline.

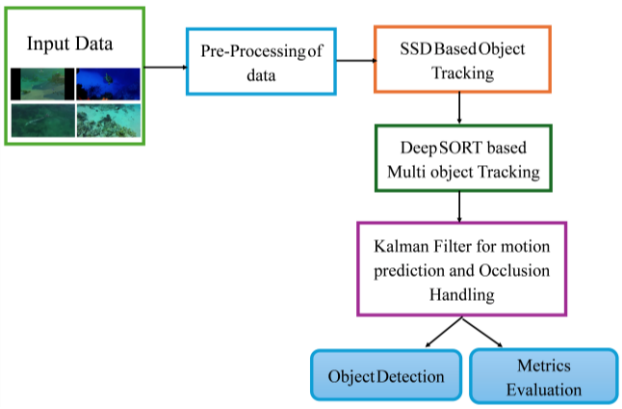


Fig. 1. Framework of proposed model.

A. Dataset

The proposed object tracking technique is evaluated on the UOT32 dataset, a comprehensive underwater tracking benchmark comprising 32 video sequences with a total of 24,241 annotated frames. Each video averages approximately 29.15 seconds in duration and 757.53 frames per sequence [26]. Designed to advance the development of state-of-the-art tracking algorithms, the UOT32 dataset captures the diverse and challenging conditions of underwater environments, including low visibility, illumination variations, and occlusions. It serves as a standard benchmark for the objective evaluation and comparison of both existing and emerging underwater object tracking methods. Table I provides the details of the

dataset. Representative sample frames from the dataset are illustrated in Fig. 2.

TABLE I. DETAILS OF UOT32 DATASET

Dataset Attributes	Description/Values
Data Type	Video sequences with continuous frames and object annotations for temporal tracking.
Annotations	Frame-by-frame annotations of object positions and identities for tracking evaluation.
Number of videos	32
Annotated frames	24,241
Averaging time	29.15 s
Frames per video	757

B. *Preprocessing of Data*

All videos were sampled at 30 fps, and individual frames were extracted for object detection and tracking. Preprocessing included color correction, contrast enhancement, and normalization to minimize the effects of underwater scattering and color attenuation. Contrast Limited Adaptive Histogram Equalization (CLAHE) and white balancing were employed to enhance the perceptual quality of frames before detection. CLAHE is a powerful image enhancement technique widely used in underwater image preprocessing to overcome poor visibility, color distortion, and non-uniform illumination caused by light scattering and absorption in water. Each frame was resized to 300×300 pixels, consistent with SSD input dimensions. The steps performed in this stage are shown in Fig. 3.

Fig. 2. Samples images of UOT32 database.

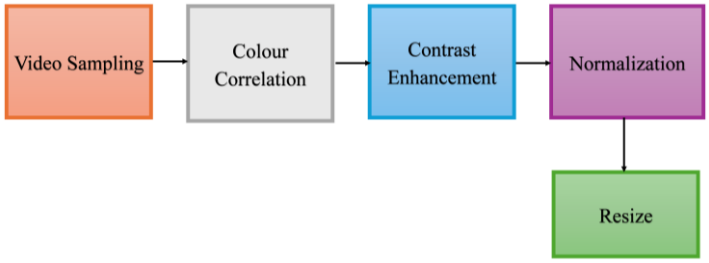


Fig. 3. Process flow of pre-processing.

C. *SSD-Based Object Detection*

One fully convolutional neural network that is used in the detection phase is the Single Shot MultiBox Detector (SSD). This SSD provides efficient detection. It can predict both the categories of objects and their bounding boxes at the same time. To adjust to underwater lighting and texturing, the SSD network employs a MobileNetV2 backbone that was pretrained on ImageNet and fine-tuned on the UOT32 dataset. The blocks in SSD-MobileNet v2 are shown in Fig. 4.

Fig. 4. SSD-MobileNet V2 blocks.

The MobileNet V2 is a lightweight depth-wise separable convolutional network that introduces inverted residuals and linear bottlenecks for efficient feature representation. The layers utilized in the work is shown in Table II.

Here, *IRB* is an inverted residual block, "*t*" is the

expansion factor, and "*c*" is the number of output channels. The 38×38 and 19×19 feature maps are typically chosen for detecting small and medium-sized underwater objects.

Training involves detecting both tiny and large underwater targets using a multi-scale feature pyramid. Smooth L1 for localization and Softmax cross-entropy for confidence are the weighted sum of the loss function. The MobileNet V2 backbone is enhanced with numerous additional convolutional layers to enhance item detection at different scales. These layers enable the network to recognize both big and small objects in a single pass by progressively reducing spatial resolution. The additional convolutional feature layers utilized are shown in Table III.

Following inference, the network produces confidence ratings and bounding boxes. To eliminate duplicate detections, it employs Non-Maximum Suppression (NMS) with an IoU threshold of 0.45. The DeepSORT tracking step takes the detected objects as input.

TABLE II. KEY LAYERS OF MOBILENET V2

Stage	Type of Layer	Kernel	Output Feature Map	Function of Layer
Conv1	Standard Conv	(3×3)/2	150×150×32	Initial Convolution
Bottleneck 1	*IRB*×1	t = 1, c = 6, s = 1	150×150×16	Low level features
Bottleneck 2	*IRB*×2	t = 6, c = 24, s = 1	75×75×24	Shallow texture features
Bottleneck 3	*IRB*×3	t = 6, c = 32, s = 1	38×38×32	First decision layer
Bottleneck 4	*IRB*×4	t = 6, c = 64, s = 1	19×19×64	Midlevel semantics
Bottleneck 5	*IRB*×3	t = 6, c = 96, s = 1	19×19×96	Stable Feature region
Bottleneck 6	*IRB*×3	t = 6, c = 160, s = 2	10×10×160	Deep Abstraction
Bottleneck 7	*IRB*×1	t = 6, c = 320, s = 1	10×10×320	Final Backbone output
Conv (1×1)	(1×1)/1	(10×10)×1280	Global Representation	

TABLE III. ADDITIONAL CONVOLUTIONAL FEATURE LAYERS

Name of the Layer	Kernel	Output Feature Map	Function of Layer
Conv8_1, Conv8_2	1×1, 3×3/2	10×10×512	Transition from backbone
Conv9_1, Conv9_2	×1, 3×3/2	5×5×256	Large object features
Conv10_1, Conv10_2	×1, 3×3/1	3×3×256	Compact Contextual cues
Conv11_1, Conv11_2	×1, 3×3/1	1×1×256	Global detection context

D. DeepSORT-Based Multi-Object Tracking

By adding a deep appearance embedding network to manage object Re-ID across frames, the DeepSORT method expands upon the conventional SORT architecture. The DeepSORT supports temporal association. To create a 128-dimensional appearance vector that represents the target's visual characteristics, each detected item is first cropped and then fed into a Convolutional Neural Network (CNN). The similarity between the currently detected items and the objects that have been tracked in the past may be measured using a cost matrix that is computed using these embeddings in conjunction with geographical information. A weighting factor (λ = 0.6) is used to optimize the assignment using the Hungarian method, which considers both appearance similarity and motion distance. At least three consecutive matches are required to validate a track, and mismatched detections start new track hypotheses.

E. Kalman Filter for Motion Prediction and Occlusion Handling

A linear Kalman Filter is used to model each active track to handle the transient object loss that might occur due to occlusions, visibility deterioration, or frame dropouts. The Kalman Filter enables motion prediction during temporary visibility loss. Here is the definition of the state vector:

$$x = \left[u, v, s, r, \dot{u}, \dot{v}, \dot{s}, \dot{r}\right]^T \tag{1}$$

where (u, v) are the centroid coordinates, s denotes the object scale, r is the aspect ratio, and the remaining terms represent their velocities.

The filter predicts the next stage using the linear motion model and is given as Eq. (2),

$$x_{t|t-1} = F \cdot x_{t-1|t-1} + w_t \tag{2}$$

where F is the state transition matrix and w_t is the process noise.

After receiving new detections from SSD, the Kalman filter performs an update process and is given as Eq. (3),

$$x_{t|t} = x_{t|t-1} + K_t\left(z_t - H \cdot x_{t|t-1}\right) \tag{3}$$

where, K_t is the Kalman gain, H is the model of observation, and z_t is the new measurement for detection. This will allow to track the combine prediction and also a good measure for estimating the state of the object accurately in conditions where noise is identified or any detection missing partially.

The constant velocity motion model predicts the object's future position when it is temporarily undetected. The Kalman filter performs prediction and correction based on incoming detections, using the Mahalanobis distance for gating to reject unlikely associations. To decide the matches of possible detections, DeepSORT uses Mahalanobis Distance (MD) as a gating criterion. MD between the predicted state $x_{t|t-1}$ and a detection z_t is given as Eq. (4).

$$d^2 = \left(z_t - H \cdot x_{t|t-1}\right)^T \left(S_T\right)^{-1} \left(z_t - H \cdot x_{t|t-1}\right) \tag{4}$$

where, S_t is the innovation covariance matrix of the Kalman Filter. After taking the covariance in the prediction and measurement into consideration, this distance estimates the distance between a detection and the expected position. If a track remains unmatched for more than 30 frames, it is terminated. This predictive mechanism enables robust occlusion recovery and minimizes identity switches.

IV. Experimental Evaluation

The proposed SSD-DeepSORT-Kalman Filter framework was implemented using matlab with a version of 2024a and a deep learning library. Experiments were conducted on a workstation equipped with an Intel Core i9-13900K CPU @ 3.00 GHz, 32 GB RAM, and an NVIDIA RTX 4090 GPU (24 GB VRAM) running on Windows 11 (64-bit).

The UOT32 underwater tracking dataset [26] was used for model training and evaluation. All videos were sampled at 30 fps, and frames were resized to 300×300 pixels to match the SSD input dimensions. During training, the learning rate was initialized at 1×10^{-4}, with the AdamW optimizer and a batch size of 16. The model was trained for 120 epochs using a multi-task loss function combining localization and classification losses. For evaluation, the framework was compared against existing GMM-Deep Sort underwater trackers in terms of precision, recall, and F1-Score.

(a) (b) (c)

Fig. 5. Output detection for given input. (a) Input; (b) Using GMM frame 150; (c) Using SDD frame 150.

Instead of using only heuristic selection, the parameter values, such as appearance-motion weighting, Kalman noise covariances, and track management thresholds, were selected based on empirical stability and resilience underwater (Table IV). Analytical derivation of optimum parameters in underwater environments is challenging because of non-stationary noise, fluctuating visibility, and unpredictable motion patterns. To guarantee consistent tracking behavior

under occlusion and detection dropouts, settings were adjusted through controlled tests on the UOT32 dataset.

TABLE IV. PARAMETERS UTILIZED

Method	Search Range	Selected Value
Appearance motion weight (λ)	0.3–0.8	0.6
Kalman process noise (Q)	10^{-4}–10^{-2}	0.001
Measurement Noise (R)	10^{-3}–10^{-1}	0.01
IoU threshold	0.3–0.6	0.45
Track Termination frames	15–40	30 frames

The experimental evaluation is conducted on the suggested benchmark dataset using the proposed methodology. The outputs achieved for the given input using GMM-Deep Sort and SSD-Deep Sort-KF are shown in Fig. 5.

The occlusion-scene evaluation is performed by conducting additional experiments by isolating video sequences from the UOT32 dataset that contain partial and full occlusion events. Occlusion frames were manually annotated based on the following criteria:

- Partial occlusion: ≥ 30% of the object area temporarily obscured.
- Severe occlusion: ≥ 60% of the object area occluded or temporarily invisible.
- Reappearance cases: Object exits and re-enters the field of view.

The proposed occlusion-aware underwater object tracking framework was evaluated using three quantitative metrics: localization error, center distance, and Intersection over Union (IoU) per frame. Together, these measures assess the UOT32 dataset's detection accuracy, spatial precision, and stability of temporal tracking across frames.

- The IoU metric is commonly used to measure the extent to which the anticipated and ground truth bounding boxes overlap in each frame. The accuracy of the detector's object localization is assessed.
- Using the centroid of the ground truth bounding box and the centroid of the predicted bounding box, the center distance can be calculated.
- The disparity in location and scale between the expected and ground truth bounding boxes is measured by the Localization Error.

The evaluation results of the above metrics for different inputs using GMM-Deep Sort and the proposed model are shown in Fig. 6.

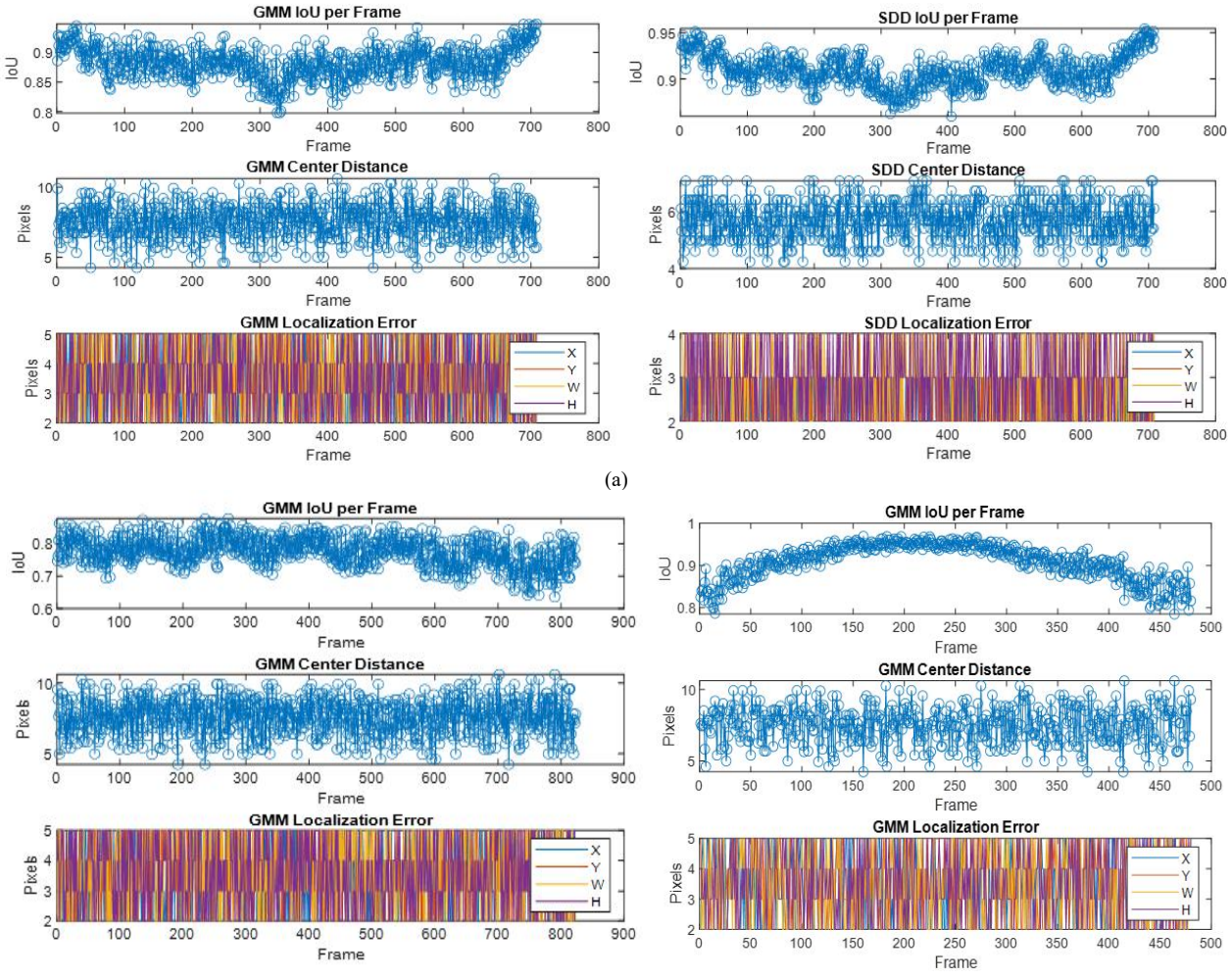
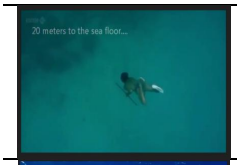
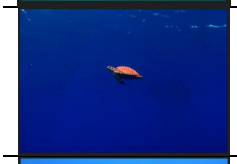


Fig. 6. Metric graphs for two different inputs: (a) GMM-Deep sort and SSD-DS-KF for input 1; (b) GMM-Deep sort and SSD-DS-KF for input 2.

TABLE V. EVALUATION METRICS UNDER OCCLUSION

Input Image	GMM-Deep SORT			SSD-Deep SORT- KF		
	Pr (%)	Re (%)	F1-Score	Pr (%)	Re (%)	F1-Score
	97.3	100	0.98	98.19	100	0.991
	98.3	100	0.99	98.68	100	0.993
	97.9	100	0.98	98.7	100	0.993
	97.2	100	0.986	98.4	100	0.992

The metrics evaluated are Precision, Recall, and F1-Score. The evaluation is performed using Eqs. (5)–(7). The achieved values are shown in Table V.

$$Precision\left(Pr\right)=\frac{TP}{TP+FP} \tag{5}$$

$$Recall\left(Re\right)=\frac{TP}{TN+FN} \tag{6}$$

$$F1-Score=2\times\frac{\left(Re\times Pr\right)}{\left(Re+Pr\right)} \tag{7}$$

where TP is True Positive; TN is True Negative; FP is False Positive; FN is False Negative.

Quantitative analysis: Three occlusion-specific performance indicators were evaluated to strengthen the proposed model. They are Occlusion Recovery Rate (ORR), Post-Occlusion Re-Identification Success Rate (POR), and Recovery Delay (RD).

ORR: Evaluated the capability of the tracker to successfully re-establish a valid trajectory after an occlusion event, and is given as Eq. (8).

$$ORR=\frac{N_{rec}}{N_{occ}}\times100 \tag{8}$$

where, N_{rec} denotes the number of occluded targets correctly recovered with preserved trajectory continuity, and N_{occ} represents the total number of annotated occlusion events.

POR: Measures identity preservation after target reappearance and is given as Eq. (9).

$$POR=\frac{N_{correctID}}{N_{Reaapearance}}\times100 \tag{9}$$

RD: Quantifies the temporal latency required for the tracker to re-lock onto a target after visibility restoration and is given as Eq. (10).

$$RD=\frac{1}{N}\sum_{i=1}^{N}\left(F_{recover}^{(i)}-F_{visible}^{(i)}\right) \tag{10}$$

where $F_{visible}$ denotes the first frame in which the object becomes visible again and $F_{recover}$ represents the frame where correct tracking identity is restored.

The results shown in Table VI are evaluated across the UOT32 dataset, a total of 142 partial and 67 severe occlusion instances were analyzed, covering short-term disappearance, overlap-based occlusion, and out-of-view reappearance scenarios. The results demonstrate that the proposed Kalman prediction combined with appearance-based association significantly reduces identity loss and enables rapid trajectory restoration following partial and severe underwater occlusions.

TABLE VI. QUANTITATIVE METRIC EVALUATION

Method	ORR (%)	POR (%)	RD (frames)
GMM Deep SORT	83.4	81.2	10.7
Proposed SSD-Deep SORT-KF	96.8	95.6	3.2

V. ABLATION STUDY

To evaluate the contribution of individual components in the proposed SSD–DeepSORT–Kalman Filter framework, an ablation study was performed. Four

different configurations were designed by progressively adding modules to the baseline SSD detector. Each model was evaluated on the UOT32 underwater object tracking dataset under identical training and testing conditions.

(1) Baseline: SSD with MobileNetV2 backbone for object detection only.
(2) SSD + DeepSORT: Added appearance embedding and data association for object tracking.
(3) DeepSORT + Kalman Filter: Included Kalman motion prediction for trajectory continuity.
(4) SSD + DeepSORT + KF + Occlusion Module: Integrated Mahalanobis distance–based gating and appearance re-identification for occlusion recovery.

Each configuration was evaluated based on Precision, Recall, F1-Score to measure overall tracking quality and is shown in Table VII.

TABLE VII. METRIC USING DIFFERENT COMBINATION OF PROPOSED MODEL

Model	Precision (%)	Recall (%)	F1-Score
SSD (Base)	91.2	93	0.918
SSD+ Deep SORT	93.12	94.6	0.934
Deep SORT+ KF	96.8	97.8	0.977
Proposed (SSD+ Deep SORT +KF+ Occlusion)	98.7	100	0.993

The ablation analysis confirms that combining spatial detection (SSD), temporal tracking (DeepSORT), and probabilistic motion modeling (Kalman Filter) results in a comprehensive, occlusion-aware framework capable of maintaining robust underwater object identities across challenging scenarios.

VI. CONCLUSION

By integrating SSD, DeepSORT, and Kalman filtering, this study suggested an underwater object tracking framework that is occlusion aware. The goal was to enhance detection accuracy and temporal consistency in demanding underwater environments. The feature extraction was made efficient by the MobileNetV2-based SSD, while color and contrast were improved by preprocessing using CLAHE and white balancing. DeepSORT used appearance embeddings to guarantee strong data association, while the Kalman Filter used Mahalanobis-distance-based gating to keep the trajectory continuous. An impressive level of accuracy was shown during evaluation on the UOT32 dataset, with an average IoU exceeding 0.75, a center distance falling below 6 pixels, and a localization error falling below 0.03. The ablation study demonstrated that the F1-Score was 7% higher after incorporating the temporal and occlusion modules than it had been with the baseline SSD. Marine monitoring, AUV navigation, and underwater surveillance are all areas that might benefit from the suggested system's occlusion-resilient tracking capabilities.

To improve generalization across different undersea habitats, future work will concentrate on combining Transformer-based feature fusion with unsupervised domain adaptation. Dense multi-object identity tracking situations are the main application for standard MOT metrics like MOTA and IDF1. In future, MOT-oriented evaluation can be performed.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Venkata Krishna Chaitanya Putrevu conducted the research work, collected the data, and wrote the paper. Dr. Chandra Bhushana Rao Kota supervised the work; all approved the final version.

REFERENCES

[1] J. Chen and M. J. Er, "Dynamic YOLO for small underwater object detection," *Artif. Intell.*, vol. 57, no. 7, 165, 2024. https://doi.org/10.1007/s10462-024-10788-1

[2] Y. Zu, L. Zhang, S. Li *et al*., "EF-UODA: Underwater object detection based on enhanced feature," *J. Mar. Sci. Eng.*, vol. 12, no. 5, 729, 2024. https://doi.org/10.3390/jmse12050729

[3] L. Chen, X. Dong, Y. Xie *et al*., "WaterPairs: A paired dataset for underwater image enhancement and underwater object detection," *Intell. Mar. Technol. Syst.*, vol. 2, no. 1, 6, 2024. https://doi.org/10.1007/s44295-024-00021-8

[4] J. Li, C. Xue, X. Luo *et al*., "Robust underwater object tracking with image enhancement and two-step feature compression," *Complex Intell. Syst*., vol. 11, no. 2, 154, 2025. https://doi.org/10.1007/s40747-024-01755-y

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition*, 2014, pp 580–587.

[6] R. Girshick, "Fast R-CNN," in *Proc. 2015 IEEE International Conf. on Computer Vision, Santiago*, 2015, pp 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2016. https:// doi. org/ 10. 1109/ TPAMI. 2016. 25770 31

[8] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv Preprint, arXiv: 2004.10934, 2020. https://doi.org/10.48550/arXiv.2004.10934

[9] W. Liu, D. Anguelov, D. Erhan *et al*., "SSD: Single shot multibox detector," in *Proc. 14th European Conf. on Computer Vision (ECCV)*, 2016, pp 21–37.

[10] H. Yang, G. Xu, S. Yi, and Y. Li, "A new cooperative deep learning method for underwater acoustic target recognition," in *Proc. OCEANS 2019-Marseill*e, 2019, pp 1–4. https://ieeexplore.ieee.org/document/8867490

[11] W. H. Lin, J. X. Zhong, S. Liu *et al*., "ROIMIX: Proposal-fusion among multiple images for underwater object detection," in *Proc. 2020 IEEE International Conf. on Acoustics, Speech and Signal Processing*, 2020, pp. 2588–2592.

[12] P. Y. Lau and S. C. Lai, "Localizing fish in highly turbid underwater images," in *Proc. International Workshop on Advanced Imaging Technology (IWAIT)*, 2021, pp 294–299.

[13] H. Ge, Y. Dai, Z. Zhu, and R. Liu, "A deep learning model applied to optical image target detection and recognition for the identification of underwater biostructures," *Machines*, vol. 10, no. 9, 809, 2022.

[14] H. Ge, Y. Dai, Z. Zhu, and X. Zang, "Single-stage underwater target detection based on feature anchor frame double optimization network," *Sensors*, vol. 22, no. 20, 7875, 2022.

[15] X. Hu, Y. Liu, Z. Zhao *et al*., "Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved YOLO-V4 network," *Comput. Electron. Agric*., vol. 185, 106135, 2021.

[16] X. Yuan, L. Guo, C. Luo *et al*., "A survey of target detection and recognition methods in underwater turbid areas," *Appl. Sci*., vol. 12, no. 10, 4898, 2022.

[17] M. Zhang, S. Xu, W. Song *et al*., "Lightweight underwater object detection based on YOLO v4 and multi-scale attentional feature fusion," *Remote Sens*., vol. 13, no. 22, 4706, 2021.
[18] S. Chen, X. Jiang, Z. Xia *et al*., "Regional proposal based underwater object tracking," in *Proc. 2022 International Conf. on Image Processing and Media Computing (ICIPMC)*, 2022, pp 30–34.
[19] Z. Wang, J. Wang, and R. Fan, "An underwater single target tracking method using siamrpn++ based on inverted residual bottleneck block," *IEEE Access*, vol. 9, pp. 25148–25157, 2021. https://doi.org/10.1109/ACCESS.2021.3056105
[20] B. Li, W. Wu, Q. Wang *et al*., "Siamrpn++: Evolution of Siamese visual tracking with very deep networks," in *Proc. 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp .4277–4286. https://doi.org/10.1109/CVPR.2019.00441
[21] M. F. R. Lee and Y. C. Chen, "Artificial intelligence based object detection and tracking for a small underwater robot," *Processes*, vol. 11, no. 2, 312, 2023. https://doi.org/10.3390/pr11020312
[22] K. Sreekala, N. N. Raj, S. Gupta *et al*., "Deep convolutional neural network with Kalman filter based objected tracking and detection in underwater communications," *Wireless Networks*, vol. 30, no. 6, pp. 1–18, 2023. https:// doi.org/10.1007/s11276-023-03290-z
[23] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and real time tracking," *IEEE International Conf. on Image Processing (ICIP)*, 2016, pp. 3464–3468. http://dx.doi.org/10.1109/ICIP.2016.7533003
[24] L. Chen, Y. Yang, Z. Wang *et al*., "Underwater target detection lightweight algorithm based on multi-scale feature fusion," *J. Mar. Sci. Eng*., vol. 11, no. 2, 320, 2023.
[25] C. Zhang, L. Liu, G. Huang *et al*., "WebUOT-1M: Advancing deep underwater object tracking with a million-scale benchmark," *Advances in Neural Information Processing Systems*, vol. 37, pp. 50152–50167, 2024. https://arxiv.org/abs/2405.19818
[26] UOT32 (underwater object tracking) dataset. [Online], Available: https://www.kaggle.com/datasets/landrykezebou/uot32-underwater-object-tracking-dataset