frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 30 November 2022
DOI 10.3389/fmars.2022.1071618

OPEN ACCESS

EDITED BY
Hong Song,
Zhejiang University, China

REVIEWED BY
Zhuang Zhou,
Beijing Institute of Technology, Zhuhai, China
Peng Ren,
China University of Petroleum (East China), China

*CORRESPONDENCE
Chang Liu
liuchang@bistu.edu.cn

SPECIALTY SECTION
This article was submitted to Ocean Observation, a section of the journal Frontiers in Marine Science

RECEIVED 16 October 2022
ACCEPTED 10 November 2022
PUBLISHED 30 November 2022

CITATION
Hao Z, Qiu J, Zhang H, Ren G and Liu C (2022) UMOTMA: Underwater multiple object tracking with memory aggregation. *Front. Mar. Sci.* 9:1071618. doi: 10.3389/fmars.2022.1071618

UMOTMA: Underwater multiple object tracking with memory aggregation

Zhicheng Hao[1], Jun Qiu[1], Haimiao Zhang[1], Guangbo Ren[2] and Chang Liu[1]*

[1]Institute of Applied Mathematics, Beijing Information Science and Technology University, Beijing, China, [2]First Institute of Oceanography, Ministry of Natural Resources, Qingdao, Shandong, China

Underwater multi-object tracking (UMOT) is an important technology in marine animal ethology. It is affected by complex factors such as scattering, background interference, and occlusion, which makes it a challenging computer vision task. As a result, the stable continuation of trajectories among different targets has been the key to the tracking performance of UMOT tasks. To solve such challenges, we propose an underwater multi-object tracking algorithm based on memory aggregation (UMOTMA) to effectively associate multiple frames with targets. First, we propose a long short-term memory (LSTM)-based memory aggregation module (LSMAM) to enhance memory utilization between multiple frames. Next, LSMAM embeds LSTM into the transformer structure to save and aggregate features between multiple frames. Then, an underwater image enhancement module M_E is introduced to process the original underwater images, which improves the quality and visibility of the underwater images so that the model can extract better features from the images. Finally, LSMAM and M_E are integrated with a backbone network to implement the entire algorithm framework, which can fully utilize the historical information of the tracked targets. Experiments on the UMOT datasets and the underwater fish school datasets show that UMOTMA generally outperforms existing models and can maintain the stability of the target trajectory while ensuring high-quality detection. The code is available *via* Github.

KEYWORDS

artificial intelligence, underwater multiple object tracking, marine environment, long-short term memory, vision transformer

1 Introduction

Multi-object tracking (MOT) is an important research topic in computer vision and is the basis of many high-level visual semantic understanding tasks. Its primary purpose is to locate multiple targets in image sequences and simultaneously track their trajectories over time. Thus, the same target has unique identification information in the image

sequence. In recent years, MOT has made considerable progress and can be found everywhere in our lives, including autonomous driving (Grigorescu et al., 2020), robot navigation (Luo et al., 2021) and video surveillance (Sreenu and Durai, 2019).

With the powerful discriminative ability of deep neural networks and the huge amount of available training data, the performance of target detection algorithms has been dramatically improved, which makes tracking-by-detection (TBD) a mainstream paradigm in MOT (Wojke et al., 2017). This method decomposes MOT into two subtasks: (1) detection, which uses a detection network to obtain bounding boxes of multiple targets in a single frame, and (2) association, which matches the detected targets with existing trajectories through an association network. However, MOT methods of the TBD paradigm still have some problems. First, the quality of the tracking results depends mainly on the detection results, which weakens the role of the association link. Second, since the whole tracking process is divided into two parts, the tracking speed of the algorithm is generally slow, which makes it challenging to meet the requirements of many application scenarios for real-time performance.

With the maturity of multitask learning methods, the joint detection and tracking (JDT) paradigm has started to attract more attention (Wang et al., 2020; Wu et al., 2021). The JDT paradigm optimizes the two abovementioned subtasks simultaneously over a backbone network. The network can output the results of object detection and the apparent features of the object corresponding to each pixel in the feature map for tracking simultaneously. This method greatly accelerates the speed of MOT, and the frames per second (FPS) can reach the real-time requirement during online tracking. However, compared with the TBD model, the tracking performance of the JDT model is not satisfactory. The feature extractor of JDT is prone to ignore the inherent variability of target localization information and identifying information in sharing embedding learning. As a result, the tracking process of JDT copes with different scale targets and occlusion situations poorly.

Although the above-mentioned MOT methods have been greatly developed, they are all designed based on pedestrian datasets and still face challenges when applied to underwater MOT (UMOT) (Panetta et al., 2021). First, the background is constant for most pedestrian MOT datasets. However, when underwater tracking is performed, the background environment can easily change drastically due to light interference or the movement of water creatures. Second, when pedestrian MOT is performed, the characteristics of most of the tracked targets are more pronounced, and *a priori* knowledge can improve the tracking accuracy. Third, when the tracking scene is switched to underwater, it is difficult for these methods to achieve good results due to the complexity of the motion of underwater creatures and the ambiguity of their characteristics.

A new memory aggregation module is propose to enhance the ability of track algorithms to correlate objects between frames in a complex underwater environment and reduce false tracking and missed tracking due to underwater environment changes. Considering both a convolutional neural network (CNN) (Liu et al., 2022b) and a transformer, improvement of the receptive field is comprehensive with deeper model depth. In contrast, for time series data, the gain from the vast improvement of the receptive field is limited. Meanwhile, more and more experiments (Tolstikhin et al., 2021) have proved that the self-attention layer does not seem to be the reason for the excellent results of the transformer. Therefore, we propose a long short-term memory (LSTM)-based memory aggregation module for historical memory fusion, named LSMAM. The overall structure of the LSMAM module follows the transformer architecture, but an LSTM-based layer replaces the multi-head attention layer. For the characteristics of complex changes in underwater scenes, we use a bi-directional long short-term memory 2D (BiLSTM2D) network Tatsunami and Taki, 2022) as a replacement. This network structure can reduce the length of sequences and produce spatially meaningful sensory fields. In short, this paper proposes an underwater MOT algorithm based on the LSMAM called **UMOTMA**. Our model is an online end-to-end tracking network with a transformer encoding and decoding structure in the main framework. The model integrates the memory module into the tracking process to fully use the location and temporal information contained in the target history information. We introduce this model in detail in Section 3.

The contributions of our proposed UMOTMA can be summarized as follows:

- We propose LSMAM, a BiLSTM2D-based memory aggregation module that is expected to improve the correlation between multiple video frames. The module's architecture follows that of a transformer, replacing the multi-headed attention layer with BiLSTM2D to enhance the modelâ€™s ability to build long temporal sequences.
- To address the problems of blurred underwater scenes and drastic environmental changes, a new end-to-end underwater MOT algorithm called UMOTMA is proposed, which integrates the LSMAM into the tracking process to improve the stability and continuity of target trajectories. The main framework of UMOTMA adopts the transformer encoder-decoder structure. In addition, it integrates an underwater image enhancement module and a memory module into the tracking process to enhance the tracking capability of the model in the complex underwater environment.
- Extensive experiments demonstrate that our method effectively improves the performance of underwater MOT, and ablation experiments show that the memory aggregation module proposed in this paper effectively improves the tracking accuracy of the

algorithm. In addition, comparative experiments are also conducted on the MOT17 datasets to demonstrate the performance and generalization ability of the method in different scenarios.

The remainder of this paper is organized as follows. Section 2 reviews the related work, and Section 3 describes the proposed method. Section 4 presents the experimental results. Section 5 discusses the results, and Section 6 draws the conclusions.

2 Relate work

2.1 Multi-object tracking

Existing MOT work is divided into two main categories: the first is the TBD paradigm, which divides MOT into two separate tasks, i.e., detection and association. The object bounding boxes are first predicted by high-performance detectors in a video frame, and then the appearance and motion features of the target are extracted by a feature extraction module; these features are then used to perform similarity value calculations. In data association, the targets are divided into different groups, and the association problem is solved by a matching algorithm to maintain the maximum global similarity while requiring the targets to achieve the one-to-one association constraint. In 2016, (Bewley et al., 2016) proposed Simple Online and Realtime Tracking (SORT), a simple algorithm framework with a fast operation speed, which has attracted widespread attention since its introduction. However, the algorithm has poor resistance to occlusion and cannot perform longer-term stable tracking. After that, (Wojke et al., 2017) further proposed Deep SORT, which uses a more reliable association metric and association method based on SORT. It can effectively track for a long time and largely reduce identity transformation in the tracking process.

Due to the rapid update of detection algorithms, more and more methods are beginning to utilize powerful detectors to obtain higher tracking performance. The You Only Look Once (YOLO) series of algorithms (Redmon et al., 2016; Redmon and Farhadi, 2018; Wang et al., 2022) has become the most popular detector because of its simplicity, efficiency, and ease of deployment. These detectors have also been adopted in a large number of methods (Chu et al., 2021; Zhang et al., 2021a; Liang et al., 2022). Most of these methods use the detection results from a single image directly for tracking.

The second class, JDT, integrates the detection and tracking modules into a single network for multitask learning to accomplish target detection and tracking simultaneously. (Wang et al., 2020) proposed the Joint Detection and Embedding (JDE) module, which utilizes DarkNet's YOLOv3 framework by adding a re-identification (ReID) branch parallel to the detection branch. The feature vector of the center point of the positive anchor frame is extracted as the apparent feature vector of the target in the feature map output from this branch. (Zhang et al., 2021b) proposed FairMOT, which is based on JDE and which chooses to perform feature extraction at the center of the estimated object. This avoids the problem that the features extracted in a coarse anchor frame may not be aligned with the center of the target, and effectively improves the performance of the tracking algorithm.

2.2 Vision transformer-based MOT

In recent years, vision transformers have been successfully applied to image recognition and video analysis with good results; as such, many works have sought to use apply them to MOT. TrackFormer (Meinhardt et al., 2022) and MOTR (Zeng et al., 2021) input the image into a CNN backbone network first to extract features and then input the extracted features into a transformer encoder. Finally, the output of the encoder and an autoregressive tracking query are used as input to the transformer decoder to perform object detection and association simultaneously. TransCenter (Xu et al., 2021) and Transtrack (Sun et al., 2020) only use transformers as a feature extractor. Their overall structure is based on encoding–decoding to pass the tracking features and learn the aggregated embedding of each object. MeMOT (Cai et al., 2022) was designed as an online tracking algorithm that performs object detection and data association under a common framework. It is capable of linking objects after a long time span, which is realized by storing the identity embeddings of the tracked objects in a large spatiotemporal memory, and by adaptively referencing and aggregating useful information from the memory as needed. Global Tracking Transformers (GTR) (Zhou et al., 2022)global is a global MOT network structure based on transformers, which uses them to encode all target features in the input video sequence and assigns the targets to different trajectories using trajectory queries.

The above works explored the different mechanisms of representing target states as dynamic embeddings. However, compared to CNN models, transformers are not yet sufficiently mature for modeling long-term spatio-temporal observations and adaptive feature aggregation.

2.3 Underwater image enhancement

Underwater image enhancement aims to improve the quality and visibility of underwater images and facilitate the acquisition of more information from the images. Contrast limited adaptive histogram equalization (CLAHE) (Reza et al., 2004) is a traditional and fast method for image enhancement. However, this method may suffer from color distortions. Fusion methods (Ancuti et al., 2012) are another classical approach for image enhancement that consider multiple enhancement techniques to

improve the quality of underwater images. In recent years, deep learning has developed rapidly and many scholars have explored the use of neural network models to improve underwater image enhancement. To avoid the requirement of paired training data, (Zhu et al., 2017) proposed the weakly supervised underwater color correction network UCycleGAN. Based on this, (Fabbri et al., 2018) proposed the underwater generative adversarial network with gradient penalty (UGAN-GP) to deal with the underwater color distortion problem, which uses UCycleGAN to generate a paired data sets for supervised training and combines the Wasserstein GAN-GP loss function to avoid model collapse. (Li et al., 2021)underwater proposed the underwater image enhancement network Ucolor, which uses medium transmission-guided multicolor space embedding. This network enriches the diversity of feature representations by incorporating features from different color spaces into a unified structure. It achieved excellent performance in experiments in various environments.

3 Materials and methods

Given a sequence of video frames $I = \{I^0, I^1, I^2, \cdots, I^T\}$, suppose there are M trajectories in frame $t-1$ and N detection targets in frame t. The goal of MOT is to complete matching between trajectories and detection targets by constructing the associated information between them and finally getting the trajectory of each detection target in the current frame. In this paper, we propose an end-to-end tracking algorithm to learn target detection and association jointly, called UMOTMA. The overall structure of the model is shown in Figure 1, which contains four main parts: (1) an underwater image enhancement module M_E. We use CLAHE, UGAN-GP, and Ucolor to implement the enhancement module, and compare their effects on tracking through experiments (Section 4). (2) A feature extraction module M_F. We use a transformer-based encoder to extract the features of the input frames. (3) A memory aggregation module LSMAM. For the memory stored in the memory buffer, LSMAM will compress it and produce an aggregated representation. (4) A feature association and update module M_A. The aggregated representation output by LSMAM is stitched with the features output by M_F as the candidate embedding for prediction. The candidate embeddings are updated using the transformer decoder, and then the new objects and tracking objects are predicted based on the updated embeddings to get the final trajectory and detection features. Finally, the history information is updated based on the results obtained from the current frame, and tracking is continued in the next frame.

3.1 Underwater image enhancement module M_E

The underwater image enhancement module M_E takes as input the original image captured by the underwater camera. Since underwater scenes are generally turbid, the main purpose of the enhancement module is to reduce color distortion effects and improve visibility. In order to choose the optimal enhancement algorithm for MOT, this paper uses CLAHE, UGAN-GP, and Ucolor as the enhancement module, where CLAHE adopts the default parameters of OpenCV, and UGAN-GP and Ucolor adopt the network model and parameters provided by the original authors. Details of the implementation are described in Section 4.

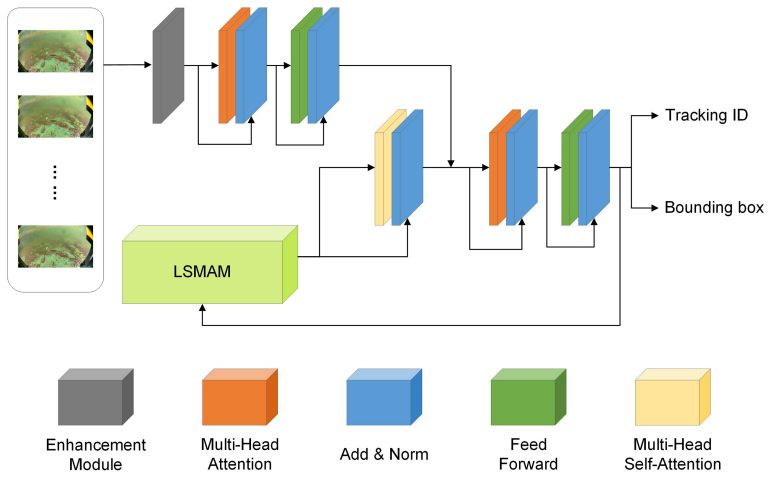

FIGURE 1
The overall structure of the UMOTMA.

3.2 Feature extraction module M_F

The M_F module is built using a transformer-based encoder for the purpose of extracting features from the images. For underwater MOT, the input to M_F is generated by the underwater image enhancement module, M_E, and for the MOT of pedestrians, the original image is fed directly to M_F. The overall framework of M_F is similar to that of MOTR (Zeng et al., 2021), where the feature map is first obtained from the input frame by a CNN, and then the feature map is fed into the transformer-based encoder, which uses the same deformable DETR (Zhu et al., 2020) structure as MOTR, and finally outputs the current frame features.

3.3 LSTM-based memory aggregation module: LSMAM

In order to reduce the length of the module and aggregate as much memory as possible while maintaining efficiency, we propose an LSTM-based memory aggregation module LSMAM, whose structure is shown in Figure 2. The overall structure of LSMAM is based on the transformer structure, in which the self-attentive layer is replaced by an LSTM-based layer called BiLSTM (Graves and Schmidhuber, 2005). In addition, we referred to the literature (Tatsunami and Taki, 2022) to improve the BiLSTM, finally deciding to using a structure similar to the vision permutator (ViP) (Hou et al., 2022), which reduces the length of the sequence, improves the accuracy and efficiency, and produces spatially meaningful sensory fields. BiLSTM consists of two layers that are replaced by combining spatial information with memory-saving memory parameters to reduce the memory cost by mixing the LSTM with memory-saving parameters. The output process is shown as follows

$$\vec{h}_{\text{for}} = \text{LSTM}_{\text{for}}(\vec{x}), \tag{1}$$

$$\overleftarrow{h}_{\text{back}} = \underset{\text{back}}{\text{LSTM}}(\overleftarrow{x}), \tag{2}$$

$$\vec{h}_{\text{back}} = \text{rearrange}\,(\overleftarrow{h}_{\text{back}}), \tag{3}$$

$$h = \text{concat}\,(\vec{h}_{\text{for}}, \vec{h}_{\text{back}})\,. \tag{4}$$

where $\vec{x}$ represents the input sequence, $\overleftarrow{x}$ represents the corresponding reverse sequence, and $\vec{h}_{\text{for}}$ and $\overleftarrow{h}_{\text{back}}$ are the outputs obtained by processing $\vec{x}$ and $\overleftarrow{x}$ the corresponding LSTMs, respectively. Here $\overleftarrow{h}_{\text{back}}$, are the outputs $\vec{h}_{\text{back}}$ rearranged in the original order, so $\vec{h}_{\text{back}}$ and $\vec{h}_{\text{for}}$ are oriented in the same direction, and finally the two are spliced to obtain h.

To parallelize the vertical and horizontal axes, LSMAM introduces two BiLSTMs for parallel processing in the left/right and top/bottom directions, named the horizontal BiLSTM and the vertical BiLSTM, respectively. For input $X \in R^{H\times W\times C}$, H represents the number of sequences in the vertical direction, W represents the number of sequences in the horizontal direction, and C is the channel's dimension. All sequences in the horizontal direction $X_w \in R^{H\times C}, w=0,1,2,\cdots W$ are input into the vertical BiLSTM, sharing the weights and hidden dimension D, and finally the output in the horizontal direction is obtained.

Similarly, all the sequences in the vertical direction $X_h \in R^{W\times C}, h=0,1,2,\cdots H$ are input into the horizontal BiLSTM to obtain the outputs. These processes are formulated as follows

$$\text{Output}_{hor} = \text{BiLSTM}\,(X_w), w = 0, 1, 2, \cdots, W, \tag{5}$$

$$\text{Output}_{ver} = \text{BiLSTM}\,(X_h), h = 0, 1, 2, \cdots, H\,. \tag{6}$$

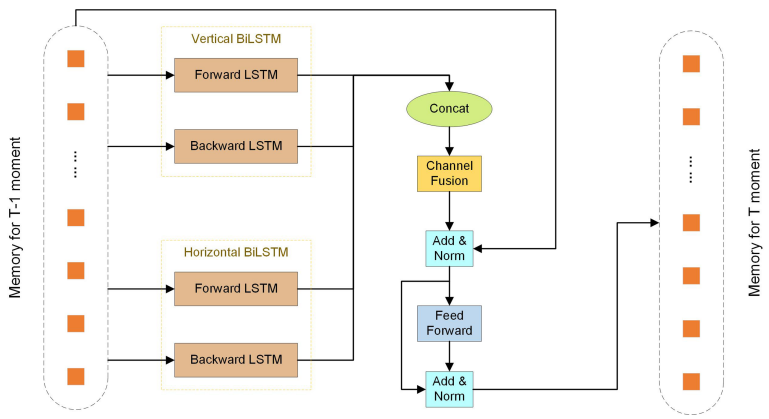

FIGURE 2
The memory aggregation module LSMAM.

We combine the horizontal and vertical results separately to obtain O^{ver} and O^{hor}, and then concatenated O^{ver} and O^{hor} to obtain the final result O

$$O = \text{concat}(O^{ver}, O^{hor}) \quad (7)$$

Note that O^{ver} and O^{hor} have the same hidden dimension $R^{W\times H\times 2C}$, which is determined by the hyperparameter of BiLSTM. Accordingly, vector O has dimensions of $R^{W\times H\times 4C}$.

3.4 Feature association and update module M_A

The overall structure of the M_A module is shown in Figure 3. It consists of multiple stacked transformer decoders, which take the image feature extracted by M_F and the output of LSMAM as the common input, where the output of LSMAM is used as the query of the decoder and the output of M_F is used as the key and value of the decoder. The decoding process produces the tracking result which contains two parts: bounding box prediction and trajectory ID prediction. For the initial frame, we generate a blank embedding to be used as historical information for feature association. In addition, to align the dimensionality of the final output, we pad each output embedding so that they can be fed to the memory aggregation module with the same dimensionality.

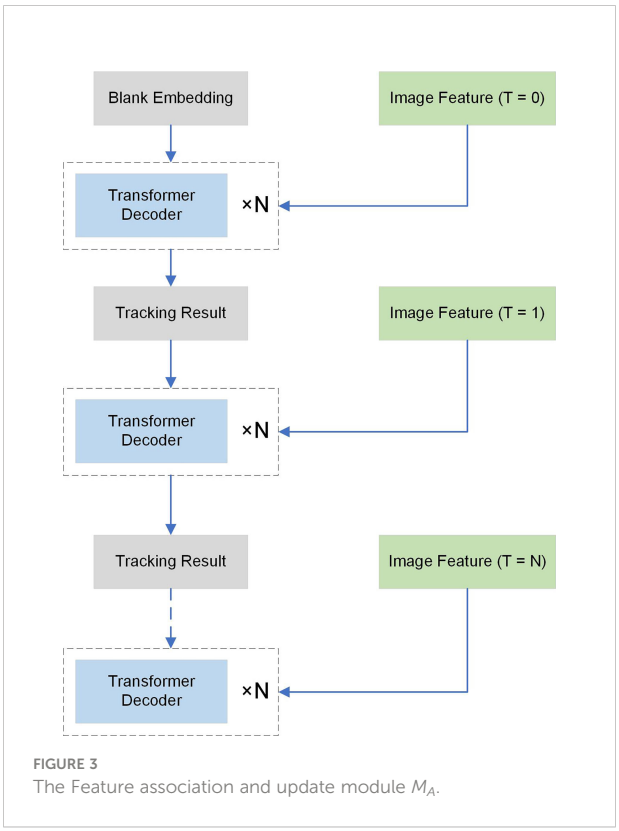

FIGURE 3
The Feature association and update module M_A.

4 Results

4.1 Datasets

To evaluate the MOT performance of UMOTMA fully, we evaluated our model on three benchmark datasets: the UMOT datasets (Zhang et al., 2020), the underwater fish school datasets (Liu et al., 2022a) and MOT17 (Milan et al., 2016). MOT17 is a representative datasets of MOT Challenge, which contains seven training subsets and seven validation subsets. All of the data are collected from the real world and labeled. The the underwater fish school datasets is a recent datasets that the images were all extracted from the observation video of a marine pasture over one year. The UMOT datasets contains four parts, which correspond to the original data set and the data set processed with CLAHE, UGAN-GP, and Fusion. Considering the size of the overall datasets, we chose to remove the Fusion data set and add the Ucolor data set instead.

4.2 Evaluation metrics

We used the same evaluation metrics as MOT Challenge to evaluate our model, where the specific metrics used include the high-order tracking accuracy (HOTA), MOT accuracy (MOTA), identity switching (IDs), recognition score (IDF1), false positives (FP), and false negatives (FN). Among them, MOTA is the most widely used metric and can closely represent human visual assessment; a better MOTA indicates that the proposed method has the ability to balance various factors. HOTA comprehensively evaluates the performance of detection and data association. IDF1 focuses more on association performance, where a higher IDF1 score indicates that the images of an object are mostly mapped to the same identity. FP and FN are defined, respectively, as the number of incorrect targets and the number of missed correct targets. It should be noted that since HOTA is a recently proposed evaluation metric, some authors have not provided this metric during their comparisons.

4.3 Implementation details

We implemented our proposed method in PyTorch 1.11. Our model was trained from scratch with a computer running Ubuntu 20.04 LTS. The entire training process was deployed on two NVIDIA RTX 3090 GPUs with memory of 48 GB. The gradient optimization method was AdamW with batch size of 12. All learning rates were initialized to 2×10^{-4} and decreased to 2×10^{-5} during the training epochs. The GFLOPs of the model is 53.1×10^{6} and the Parameters is 2×10^{11}. The model was initially trained on the UMOT datasets and the underwater fish school datasets with 100 epochs, and then fine-tuned using 60 epochs. It

took 48 hours in total. The initial training on MOT17 consisted of 60 epochs, and fine-tuning used 40 epochs. The total training time was about 36 hours. The fine-tuning started with an initial learning rate, which decreased after 10 epochs. Depending on the nature of the trajectory tracking, the total number of tracks per frame varied. In order to stack the results of multiple frames into a batch, we complemented each frame result with blank trace results to align the lengths of the trace results of all frames.

4.4 Comparison with state-of-the-art methods on the UMOT datasets

Table 1 shows the results of the method proposed in this paper relative to the other tracking methods, including DeepSORT (Wojke et al., 2017), CenterTrack (Zhou et al., 2020), TrackFormer (Meinhardt et al., 2022), GSDT (Wang et al., 2021), and MOTR (Zeng et al., 2021) on the UMOT datasets. Since some of these methods have not been previously applied to the UMOT datasets; so, to be fair, we re-implemented these methods on the experimental equipment and obtained their results for the UMOT datasets. Each metric has an arrow next to it, with "↑" indicating that the higher the metric is, the better, and "↓" indicating that the lower it is, the better.

As can be seen in Table 1, UMOTMA achieves excellent results of 52.3% and 61.1% in terms of the MOTA and IDF1 metrics for the UMOT datasets for underwater scenes, 8.2% and 7.7% better than the second-placed method, respectively. In addition to MOTA and IDF1, other metrics applied to UMOTMA also reflect some improvements relative to the other methods. However, our proposed method does not perform the best in terms of FP and IDs, probably because the extracted features of the false detection targets are very similar to those of the correct targets, which in turn lead to false detections and incorrect associations. The excellent results of MOTA and IDF1 show that our model has good tracking performance and can maintain a stable continuation of the tracking trajectory. This is mainly due to our incorporation of the memory aggregation module, which enables feature extraction of historical information through aggregation of past frame tracking results and improves the accuracy of the associated trajectories with current frame targets. Figure 4 shows some of the results of our underwater tracking.

4.5 Comparison with state-of-the-art methods on the underwater fish school datasets

Recently, the underwater fish school datasets is introduced to provides a better choice to verify the underwater multi-object tracking performance. We further conduct the experiments on the underwater fish school datasets and perform the performance comparison with state-of-the-art methods in Table 2. It shows that UMOTMA achieves much better performance on the underwater fish school datasets. Our method gets a much higher MOTA score, surpassing JDE by 2.3%. For the IDF1 metric, our method also achieves much better performance than JDE (81.1% vs. 72.2%). While for the IDs metric, UMOTMA is inferior to some state-of-the-art methods. It means that UMOTMA performs well on temporal motion learning while the tracking performance is not that stable. The large improvements on MOTA are mainly from the memory aggregation network.

4.6 Ablation studies

4.6.1 UMOTMA components

In this subsection, we research the effectiveness of the different components in UMOTMA, including the underwater enhancement module (M_E), and the memory aggregation module (LSMAM). The experiment results are shown in Table 3. Baseline represents only using the feature extraction module (M_F), and association module (M_A), without using $M_E$and LSMAM, whose results are relatively poor. Baseline + $M_E$means the underwater image enhancement module has been added to the baseline model. The underwater image enhancement method can improve tracking performance by increasing image visibility, so the tracking effect is effectively improved by adding M_E, as

TABLE 1 Comparison of the methods on the UMOT test set.

Methods	IDF1↑	MOTA↑	IDs↓	FP↓	FN↓
DeepSORT	44.7	26.2	18	**312**	2837
GSDT	52.2	39.4	30	430	3045
CenterTrack	50.6	42.4	**15**	914	3101
TrackFormer	53.4	44.1	34	583	3294
MOTR	52.7	42.6	18	483	3516
UMOTMA	**61.1**	**52.3**	28	499	**2809**

In each column, the best result is in bold.

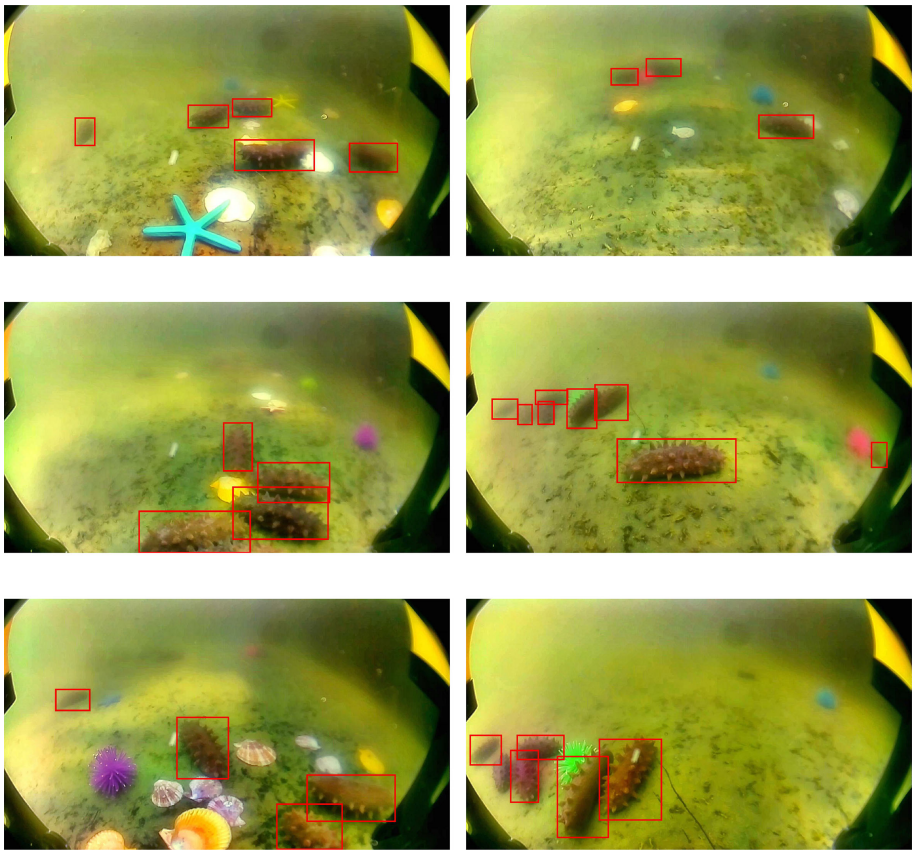

FIGURE 4
Some examples of tracking results produced by our proposed UMOTMA.

indicated by the MOTA index, which has increased by 2.2%. However, due to the increased image visibility, the information captured by the model during feature extraction increases substantially, resulting in a surge of IDs during the association process, which affects the tracking stability of the model. UMOTMA indicates the further addition of the LSMAM to Baseline + M_E. In terms of metrics, the addition of LSMAM increases the MOTA metric by 3.8%, IDF1 by 5.9%, and ID by a factor of 7. The reason it achieves such good results is that LSMAM allows the model to use more historical information to match trajectories with current frame features, effectively reducing the fluctuations of tracking trajectories due to ID, reducing the IDs to a certain extent, and improving the tracking effect.

TABLE 2 Comparison of the methods on the underwater fish school test set.

Methods	IDF1↑	MOTA↑ ‘	IDs↓
SORT	75.4	71.6	342
DeepSORT	77.4	75.4	301
JDE	72.2	76.1	453
Bytetrack	79.3	77.6	**249**
UMOTMA	**81.1**	**78.4**	277

In each column, the best result is in bold.

4.6.2 The influence of different LSTMs on tracking

LSTM is a classical neural network, and there are many variants based on it. To investigate the effect of different LSTMs on the model tracking effect, we used BiLSTM and BiLSTM2D to replace LSTM in the LSMAM module separately. The results of the ablation experiments are shown in Table 4. Compared with the baseline LSTM only, the tracking accuracy of the network with both the BiLSTM and BiLSTM2D structures has significantly improved. The MOTA metric increases by 4.8% and 7.0%, respectively, and the IDF1 metric increases by 3.1% and 8.7%, respectively. The FP and FN metrics also improve to different degrees. These results show that the memory module

TABLE 3 Ablation study on UMOT datasets.

Methods	IDF1↑	MOTA↑	IDs↓	FP↓	FN↓
Baseline	54.7	46.3	**10**	577	3170
Baseline + M_E	55.2	48.5	35	532	3035
UMOTMA	**61.1**	**52.3**	28	**499**	**2809**

In each column, the best result is in bold.

with BiLSTM2D has a significant improvement on the tracking accuracy of the model, and by using a bi-directional 2D structure, the module can perform better aggregation of the information contained in the time series data.

4.6.3 Comparison of underwater image enhancement modules

Underwater image enhancement is a fundamental task in the field of computer vision, and many excellent works have subsequently emerged (Reza et al., 2004; Ancuti et al., 2012; Fabbri et al., 2018; Li et al., 2021). In order to study the enhancement effect of enhancement methods for underwater MOT, we used different algorithms to build enhancement modules and evaluate the amount of enhancement imparted by the different modules *via* experiments. The experimental results are shown in Table 5. From Table 5, we can see that all three image enhancement algorithms have a certain degree of improvement in terms of the MOTA metric, but the IDs metric indicates there are different degrees of degradation, among which UGAN-GP causes the most serious amount of degradation. Figure 5 shows the images produced by the different enhancement modules. Through a comparison of the images, we can find that the image contrast improvement brought by the UGAN-GP algorithm is the highest, which directly changes the color of the image background, thus causing a significant decrease in the ID index. The Ucolor processed image is more consistent with the original image in terms of color, and therefore produced the best IDs metric value.

4.7 The performance on pedestrian datasets

To evaluate the tracking capability of our method UMOTMA in different scenarios, we conducted experiments on the MOT17 datasets and compared the results with those obtained with the other tracking methods. Table 6 lists the metric results of our proposed method UMOTMA against those of the other state-of-the-art MOT methods. It can be seen that UMOTMA obtains competitive tracking accuracy on the MOT17 datasets, achieving the best results for the HOTA, MOTA, IDF1, and FN metrics, and the second-best IDs metric value.

In general, the experimental results for the datasets in the two scenarios show that the method proposed in this paper has obvious advantages in terms of comprehensive performance and tracking accuracy, and performs well in different scenarios. In particular, the results achieved by UMOTMA in the underwater scenario are significantly higher than those of the other existing methods.

5 Discussion

The main purpose of MOT is to assign IDs to detected targets and keep the IDs of the same targets unchanged in the subsequent frames. Most previous works were performed based on pedestrian datasets. However, for underwater scenes, unfavorable conditions such as occlusion, background interference, and motion blurring appear more frequently, so it becomes extremely difficult to keep the tracking stable when performing MOT in underwater environments. In this paper, we propose a new end-to-end MOT algorithm called UMOTMA. The main advantage of UMOTMA over other methods is the introduction of a depth LSTM-based memory aggregation module (LSMAM), which enhances the model in terms of correlation features by fully aggregating the information contained in past frames to maintain stable tracking in complex environments. The experimental results on the UMOT datasets and the underwater fish school datasets showed that our proposed UMOTMA has

TABLE 4 Ablation study about different LSTM.

Methods	IDF1↑	MOTA↑	IDs↓	FP↓	FN↓
LSMAM(LSTM)	52.4	45.3	**19**	813	2999
LSMAM(BiLSTM)	55.1	49.3	26	**485**	3078
LSMAM(BiLSTM2D)	**61.1**	**52.3**	28	499	**2809**

In each column, the best result is in bold.

TABLE 5 Comparison of the different enhancement module.

Enhancement	IDF1↑	MOTA↑	IDs↓	FP↓	FN↓
Origin	51.9	46.7	**12**	484	3011
CLAHE	55.5	49.1	28	482	3060
UGAN-GP	48.3	49.6	34	**312**	2837
Ucolor	**61.1**	**52.3**	28	499	**2809**

In each column, the best result is in bold.

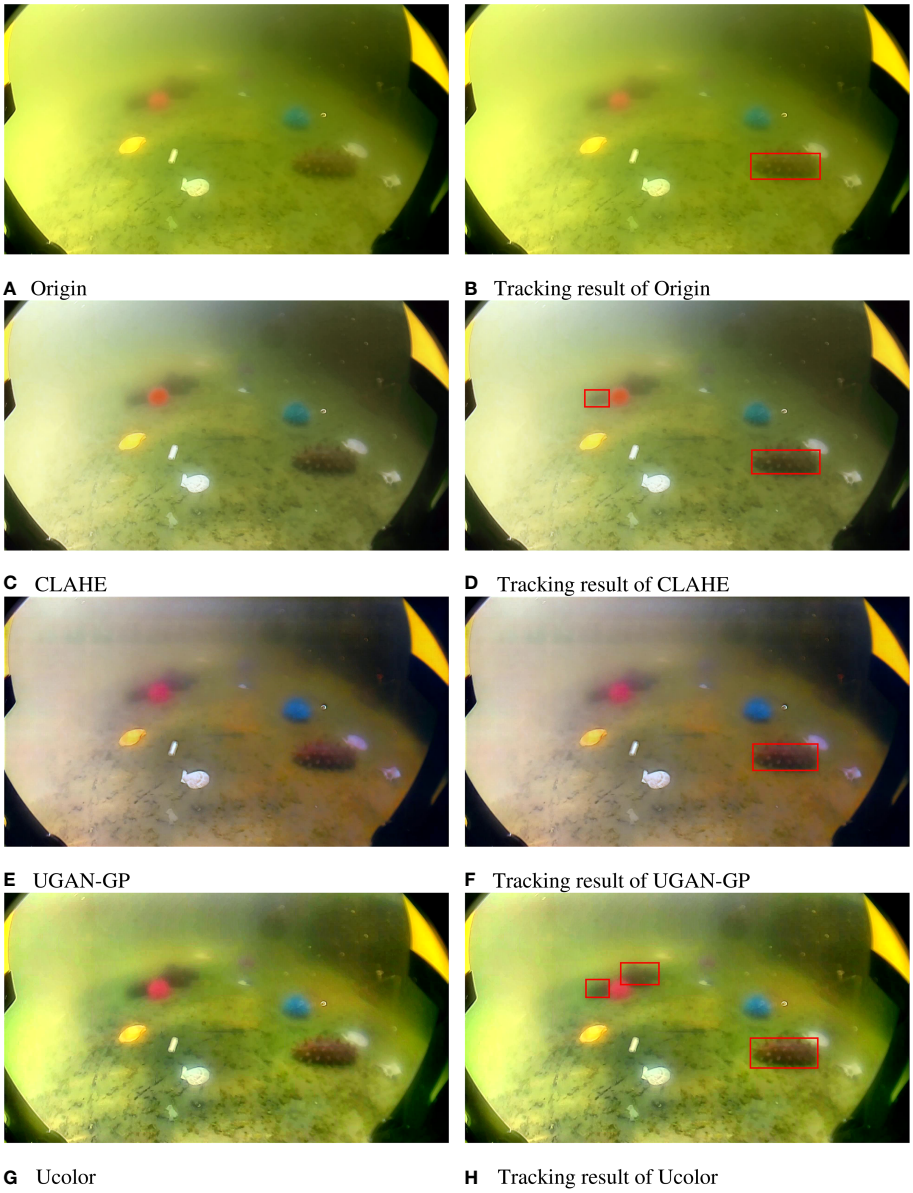

FIGURE 5
Visualization of the results generated by the different enhancement modules.

TABLE 6 Comparison of the state-of-the-art methods on the MOT17 test set.

Methods	HOTA↑	IDF1↑	MOTA↑	IDs↓	FP↓
DeepSORT	–	61.2	60.3	2442	36111
CenterTrack	52.2	64.7	67.8	3039	**18498**
TraDeS	52.7	63.9	69.1	3555	20892
GSDT	55.5	68.7	66.2	3318	26339
TrackFormer	–	63.9	65.0	3528	70443
MOTR	57.2	68.4	71.9	**2115**	32355
UMOTMA	**57.6**	**68.8**	**72.3**	2436	37149

In each column, the best result is in bold.

excellent performance in underwater MOT, and several evaluation metrics reached optimal performance, which proves the effectiveness of the proposed method.

In order to explore the roles that LSMAM performs in the tracking process further, we conducted extensive ablation experiments. The results are shown in Table 3. After adding the LSMAM module, the MOTA and IDF1 metrics of the model increased respectively by 3.8% and 5.9%, and the IDs decreased by a factor of 7, indicating that the memory aggregation module proposed in this paper effectively improves the tracking accuracy of the model. However, We also found that the underwater image enhancement module and the memory aggregation module increase the calculation volume of the model, which leads to slow inference speed of the model and makes it difficult to satisfy some underwater scenarios for real-time MOT. Therefore, it will be an important direction of subsequent work to optimize the model inference speed.

6 Conclusions

This paper proposed a novel deep LSTM-based end-to-end underwater MOT model named UMOTMA. We introduced a memory aggregation module to guide the matching association link between past frame trajectories and current frame features. In the memory aggregation module, we use LSTM for memory aggregation instead of a CNN or transformer, as employed in conventional approaches, which effectively improves the algorithmâ€™s utilization of target information in past frames. Experimental results on the UMOT datasets and the underwater fish school datasets showed that our proposed UMOTMA achieves optimal results in terms of several MOT metrics, and it was significantly better than the second-best method. The experimental results for the MOT17 datasets also showed that our method has a tracking accuracy comparable to other state-of-the-art MOT methods for pedestrian scenes. In addition, we conducted an extensive ablation study to demonstrate the contribution of each component of the proposed MOT framework to the tracking process and briefly discussed the impact of different image enhancement modules on MOT in underwater environments.

In general, our proposed method has good generality and can be adapted to both surface and underwater application scenarios. Especially in the latter, UMOTMA shows exceptionally competitive performance. In the future, we will use our model for MOT of critical underwater scenarios and exploration of marine biological activities.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/Zxl19990529/Underwater-Multiple-Object-Tracking-Dataset.

Author contributions

ZH designed the method with experiments and wrote the first draft of the manuscript supervised by CL and HZ. GR anb JQ critically reviewed the initial manuscript and provided helpful input. All authors contributed to the article and approved the submitted version.

Funding

The authors would like to acknowledge the financial support from the National Natural Science Foundation of China (under Grant Nos. 61871042, 61931003, 62171044, 12101061), the Natural Science Foundation of Beijing (Grants No. 4222004), and QinXin Talents Cultivation Program(Beijing Information Science and Technology University), also from the China High-

Resolution Earth Observation System Program under Grant 41-Y30F07-9001-20/22.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ancuti, C., Ancuti, C. O., Haber, T., and Bekaert, P. (2012). "Enhancing underwater images and videos by fusion," in *Proc. conf. comput. vis. pattern recog* (Providence, RI, USA: IEEE), 81–88.

Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). "Simple online and realtime tracking," in *Proc. int. conf. image process* (Phoenix, AZ, USA: IEEE), 3464–3468.

Cai, J., Xu, M., Li, W., Xiong, Y., Xia, W., Tu, Z., et al. (2022). "Memot: Multi-object tracking with memory," in *Proc. conf. comput. vis. pattern recog* (New Orleans, LA, USA: IEEE), 8090–8100.

Chu, P., Wang, J., You, Q., Ling, H., and Liu, Z. (2021). Transmot: Spatial-temporal graph transformer for multiple object tracking. *arXiv*. doi: 10.48550/arXiv.2104.00194

Fabbri, C., Islam, M. J., and Sattar, J. (2018). "Enhancing underwater imagery using generative adversarial networks," in *Proc. int. conf. robot. autom* (Brisbane, QLD, Australia: IEEE), 7159–7165.

Graves, A., and Schmidhuber, J. (2005). "Framewise phoneme classification with bidirectional lstm networks," in *Proc. int. joint conf. neural netw*, (Montreal, QC, Canada: IEEE) 4, 2047–2052.

Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *J. Field Robotics* 37, 362–386. doi: 10.1002/rob.21918

Hou, Q., Jiang, Z., Yuan, L., Cheng, M.-M., Yan, S., and Feng, J. (2022). Vision permutator: A permutable mlp-like architecture for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1. doi: 10.1109/TPAMI.2022.3145427

Liang, C., Zhang, Z., Zhou, X., Li, B., Zhu, S., and Hu, W. (2022). Rethinking the competition between detection and reid in multiobject tracking. *IEEE Trans. Image Process.* 31, 3182–3196. doi: 10.1109/TIP.2022.3165376

Li, C., Anwar, S., Hou, J., Cong, R., Guo, C., and Ren, W. (2021). Underwater image enhancement *via* medium transmission-guided multi-color space embedding. *IEEE Trans. Image Process.* 30, 4985–5000. doi: 10.1109/TIP.2021.3076367

Liu, T., He, S., Liu, H., Gu, Y., and Li, P. (2022a). A robust underwater multiclass fish-school tracking algorithm. *Remote Sens.* 14, 2072–4292. doi: 10.3390/rs14164106

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022b). "A convnet for the 2020s," in *Proc. conf. comput. vis. pattern recog* (New Orleans, LA, USA: IEEE), 11976–11986.

Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., and Kim, T.-K. (2021). Multiple object tracking: A literature review. *Artif. Intell.* 293, 103448. doi: 10.1016/j.artint.2020.103448

Meinhardt, T., Kirillov, A., Leal-Taixe, L., and Feichtenhofer, C. (2022). "Trackformer: Multi-object tracking with transformers," in *Proc. conf. comput. vis. pattern recog* (New Orleans, LA, USA: IEEE), 8844–8854.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K. (2016). Mot16: A benchmark for multi-object tracking. *arXiv*. doi: 10.48550/arXiv.1603.00831

Panetta, K., Kezebou, L., Oludare, V., and Agaian, S. (2021). Comprehensive underwater object tracking benchmark dataset and underwater image enhancement with gan. *IEEE J. Oceanic Eng.* 47, 59–75. doi: 10.1109/JOE.2021.3086907

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: Unified, real-time object detection," in *Proc. conf. comput. vis. pattern recog* (Las Vegas, NV, USA: IEEE), 779–788.

Redmon, J., and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv*. doi: 10.48550/arXiv.1804.02767

Reza, A. M. (2004). Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *J. VLSI Signal Process. Syst. signal image video Technol.* 38, 35–44. doi: 10.1023/B:VLSI.0000028532.53893.82

Sreenu, G., and Durai, S. (2019). Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *J. Big Data* 6, 1–27. doi: 10.1186/s40537-019-0212-5

Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., et al. (2020). Transtrack: Multiple object tracking with transformer. *arXiv*. doi: 10.48550/arXiv.2012.15460

Tatsunami, Y., and Taki, M. (2022). Sequencer: Deep lstm for image classification. *arXiv*. doi: 10.48550/arXiv.2205.01972

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., et al. (2021). "Mlp-mixer: An all-mlp architecture for vision," in *Advances in neural information processing systems*, (Montreal, QC, Canada: Curran Associates, Inc) 34, 24261–24272.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv*. doi: 10.48550/arXiv:2207.02696

Wang, Y., Kitani, K., and Weng, X. (2021). "Joint object detection and multi-object tracking with graph neural networks," in *Proc. int. conf. robot. autom* (Xi'an, China: IEEE), 13708–13715.

Wang, Z., Zheng, L., Liu, Y., Li, Y., and Wang, S. (2020). "Towards real-time multi-object tracking," in *Proc. lect. notes comput. sci* (Glasgow, UK: Springer), 107–122.

Wojke, N., Bewley, A., and Paulus, D. (2017). "Simple online and realtime tracking with a deep association metric," in *Proc. int. conf. image process* (Beijing, China: IEEE), 3645–3649.

Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., and Yuan, J. (2021). "Track to detect and segment: An online multi-object tracker," in *Proc. conf. comput. vis. pattern recog* (Nashville, TN, USA: IEEE), 12352–12361.

Xu, Y., Ban, Y., Delorme, G., Gan, C., Rus, D., and Alameda-Pineda, X. (2021). Transcenter: Transformers with dense representations for multiple-object tracking. *arXiv*. doi: 10.48550/arXiv.2103.15145

Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., and Wei, Y. (2021). Motr: End-to-end multiple-object tracking with transformer. *arXiv*. doi: 10.48550/arXiv.2105.03247

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., et al. (2021a). Bytetrack: Multi-object tracking by associating every detection box. *arXiv*. doi: 10.48550/arXiv.2110.06864

Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W. (2021b). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vision* 129, 3069–3087. doi: 10.1007/s11263-021-01513-4

Zhang, X., Zeng, H., Liu, X., Yu, Z., Zheng, H., and Zheng, B. (2020). *In situ* holothurian noncontact counting system: A general framework for holothurian counting. *IEEE Access* 8, 210041–210053. doi: 10.1109/ACCESS.2020.3038643

Zhou, X., Koltun, V., and Krähenbühl, P. (2020). "Tracking objects as points," in *Proc. lect. notes comput. sci* (Glasgow, UK: Springer), 474–490.

Zhou, X., Yin, T., Koltun, V., and Krähenbühl, P. (2022). "Global tracking transformers," in *Proc. conf. comput. vis. pattern recog* (New Orleans, LA, USA: IEEE), 8771–8780.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. conf. comput. vis. pattern recog* (Venice, Italy: IEEE), 2223–2232.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv*. doi: 10.48550/arXiv.2010.04159