

# Asynchronous Trajectory Matching-Based Multimodal Maritime Data Fusion for Vessel Traffic Surveillance in Inland Waterways

Yu Guo<sup>1</sup>, Member, IEEE, Ryan Wen Liu<sup>1</sup>, Member, IEEE, Jingxiang Qu<sup>1</sup>, Yuxu Lu<sup>1</sup>, Fenghua Zhu<sup>1</sup>, Senior Member, IEEE, and Yisheng Lv<sup>1</sup>, Senior Member, IEEE

**Abstract**—The automatic identification system (AIS) and video cameras have been widely exploited for vessel traffic surveillance in inland waterways. The AIS data could provide vessel identity and dynamic information on vessel position and movements. In contrast, the video data could describe the visual appearances of moving vessels without knowing the information on identity, position, movements, etc. To further improve vessel traffic surveillance, it becomes necessary to fuse the AIS and video data to simultaneously capture the visual features, identity, and dynamic information for the vessels of interest. However, the performance of AIS and video data fusion is susceptible to issues such as data spatial difference, message asynchronous transmission, visual object occlusion, etc. In this work, we propose a deep learning-based simple online and real-time vessel data fusion method (termed DeepSORVF). We first extract the AIS- and video-based vessel trajectories, and then propose an asynchronous trajectory matching method to fuse the AIS-based vessel information with the corresponding visual targets. In addition, by combining the AIS- and video-based movement features, we also present a prior knowledge-driven anti-occlusion method to yield accurate and robust vessel tracking results under occlusion conditions. To validate the efficacy of our DeepSORVF, we have also constructed a new benchmark dataset (termed FVessel) for vessel detection, tracking, and data fusion. It consists of many videos and the corresponding AIS data collected in various weather conditions and locations. The experimental results have demonstrated that our method is capable of guaranteeing high-reliable data fusion and anti-occlusion vessel tracking. The DeepSORVF code and FVessel dataset are publicly available at <https://github.com/gy65896/DeepSORVF> and <https://github.com/gy65896/FVessel>, respectively.

**Index Terms**—Inland waterways, vessel traffic surveillance, deep neural network, anti-occlusion tracking, data fusion.

## I. INTRODUCTION

THE autonomy and intelligence of the inland waterways surveillance system play a significant role in the development of inland waterborne transportation, which can effectively reduce the labor cost of the supervisory departments and ensure the safety of vessel navigation. To accomplish this objective, the vessel traffic service (VTS) system is capable of providing effective situational awareness by using the automatic identification system (AIS), radar, closed circuit television (CCTV) [1], etc. To further increase the capability of situational awareness, many intelligent technologies for single sensor have been presented, e.g., AIS-based vessel trajectory prediction [2], radar-based object detection [3], and video-based object detection [4]. It is well known that each type of sensor has its own advantages and disadvantages under the same scenarios. As a consequence, numerous efforts have been devoted to simultaneously exploiting the multi-source data [1], [5], [6], [7], [8], [9], [10], [11] to promote the traffic situational awareness for maritime transportation systems. However, these fusion methods mainly just take into consideration the positional relationship of the same target at a certain moment. It thus becomes difficult to guarantee high-quality data fusion, especially for the existence of time delay, missing data, random outliers, etc. The same moving vessels essentially share similar navigation behaviors, which could be represented using the time-series data, e.g., spatio-temporal trajectories. To further improve the stability and accuracy of data fusion, we will first extract the vessel trajectories from the raw sensing data, and then propose a trajectory matching-based fusion method (termed DeepSORVF) in this work.

### A. Motivation and Contribution

Owing to the remote, intuitive, and real-time advantages of CCTV, terrestrial video surveillance systems have been widely used in inland waterborne transportation to improve the ability of traffic situational awareness and vessel abnormal behavior monitoring [12]. In particular, massive monitoring

Manuscript received 26 January 2023; revised 9 April 2023 and 31 May 2023; accepted 8 June 2023. Date of publication 22 June 2023; date of current version 1 November 2023. This work was supported in part by the National Key R&D Program of China under Grant 2022YFB4300300, in part by the National Natural Science Foundation of China under Grant 52271365, and in part by the Fundamental Research Funds for the Central Universities under Grant 2023-vb-045. The Associate Editor for this article was B. Du. (Corresponding authors: Ryan Wen Liu; Fenghua Zhu.)

Yu Guo, Ryan Wen Liu, Jingxiang Qu, and Yuxu Lu are with the School of Navigation, Wuhan University of Technology, Wuhan 430063, China, and also with the State Key Laboratory of Maritime Technology and Safety, Wuhan 430063, China (e-mail: yuguo@whut.edu.cn; wenliu@whut.edu.cn; qujx@whut.edu.cn; yuxulu@whut.edu.cn).

Fenghua Zhu and Yisheng Lv are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: fenghua.zhu@ia.ac.cn; yisheng.lv@ia.ac.cn).

Digital Object Identifier 10.1109/TITS.2023.3285415

cameras can provide indispensable visual information for guaranteeing maritime safety. To fully use these visual features, many efforts have focused on the research of vessel detection and tracking to meet the requirement of intelligent supervision [13], [14], [15]. However, these methods could only detect the moving vessel from the video images. It is intractable to achieve important identity information (e.g., vessel name and size) and dynamic information (e.g., vessel speed and course).

Other maritime awareness equipment, such as AIS and radar, could provide much richer attribute information about the vessel. In particular, the AIS data contains rich vessel identity and spatio-temporal information, which makes it play an essential role in analyzing vessel abnormal behaviors. The AIS data mainly contains static and dynamic information, e.g., maritime mobile service identification (MMSI), vessel size, speed, course, position, etc. However, the AIS data essentially suffers from the inconsistency of time intervals, which limits its application in maritime intelligent transportation [16]. The radar has been widely used in near port supervision since it can provide the accurate distance and bearing of vessels. Unfortunately, some radar equipment is forbidden to be installed in populated regions to avoid the high-frequency electromagnetic radiation harming the health of people [1]. In the literature [17], many methods have been proposed to robustly and accurately fuse the AIS and radar data. As long we simultaneously collect the AIS, radar, and video data, we can directly adopt the existing advanced methods to fuse the AIS and radar data, and then implement fusion with the video data. Intuitively, the fusion of AIS and video data seems more difficult because of the different coordinate systems, asynchronous data collection, different data structures, etc. Therefore, we tend to only fuse the AIS and video data to enhance the traffic situational awareness for intelligent surveillance in inland waterways.

In this work, we propose a deep learning-based simple online and real-time vessel data fusion method (termed DeepSORVF) for promoting inland waterways surveillance. The main contributions of this work are as follows:

- We build two simple yet efficient methods to extract the AIS- and video-based vessel trajectories for data fusion, respectively. To avoid the interference of vessel occlusion, we propose a novel anti-occlusion tracking method, which adopts the historical fusion results as prior knowledge to predict the bounding boxes and video-based vessel trajectories in the occluded region.
- We design a novel asynchronous trajectory matching method to achieve the robust fusion of AIS and video data. The proposed method adopts an enhanced fast dynamic time warping algorithm for trajectory similarity measure and employs an AIS/video association method to decrease the computational cost and increase the stability.
- We construct a public benchmark dataset (termed FVessel) for vessel detection, tracking, and data fusion, which consists of many videos and the corresponding AIS data collected in various weather conditions and locations.

To the best of our knowledge, the proposed DeepSORVF is the first trajectory matching-based computational method to fuse the AIS and video data for inland waterways surveillance.

Meanwhile, we have verified the effectiveness of the proposed method on our newly-developed FVessel dataset.

## B. Organization

The rest of this paper is organized as follows. Section II briefly reviews the recent research on object detection, tracking, AIS and video data fusion. In Section III, the proposed data fusion framework is described in detail. Section IV implements extensive comprehensive experiments to demonstrate the effectiveness of our method. Finally, Section V summarizes the main contributions of this work.

## II. RELATED WORKS

This section mainly introduces the recent studies related to our work, i.e., multi-vessel detection and tracking, AIS and video data fusion.

### A. Multi-Vessel Detection and Tracking

Multi-object detection and tracking methods are generally divided into two categories, namely traditional and deep learning methods. Due to the particularity of the research issue, this section mainly reviews the related works on vessel detection and tracking.

1) *Traditional Methods*: Background subtraction (BS) is a classic object detection method. Although many BS-based methods are proposed to detect conventional objects, these methods still achieve poor precision in vessel detection [18]. To improve the accuracy of the BS-based method, Hu et al. [19] designed a robust foreground detection and background update method to effectively reduce the influence of waves. Bloisi et al. [20] proposed an independent multi-modal background subtraction (IMBS) algorithm. In particular, this algorithm models highly dynamic backgrounds (e.g., water) by creating a “discretization” of an unknown distribution. Furthermore, other types of vessel detection methods are proposed. For instance, Zhu et al. [21] designed a hierarchical complete-based vessel detection approach for spaceborne optical images. Zhang et al. [22] proposed a vessel detection algorithm using the discrete cosine transform (DCT)-based Gaussian mixture model (GMM) for efficient visual maritime surveillance on non-stationary surface platforms. Chen et al. [23] achieved the vessel object tracking using multi-view learning and sparse representation. Although many techniques have been introduced to improve the performance of detectors, hand-designed features still produce poor robustness in vessel detection. Meanwhile, the high computational complexity of some methods will hinder their practical applications.

2) *Deep Learning Methods*: With the emergence and rapid development of graphics processing units (GPU), deep learning technology is widely used in the field of image processing. Many deep learning methods are proposed for object detection, e.g., region-based convolutional neural network (R-CNN) [24], [25], single shot multibox detector (SSD) [26], [27], and you only look once network (YOLO) [28], [29], [30], [31]. Based on these object detection networks, many vessel detection methods are further researched.

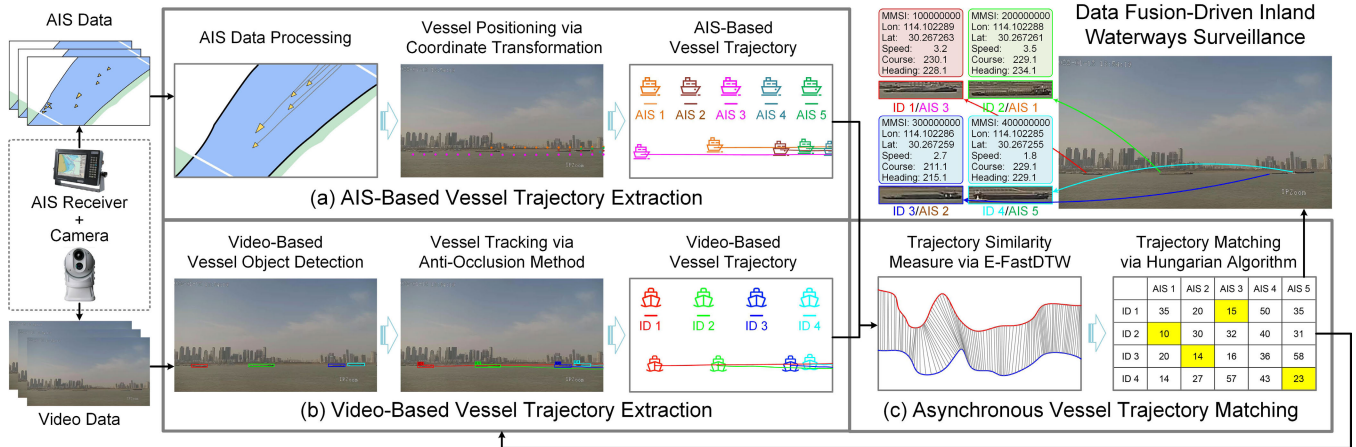


Fig. 1. The architecture of the proposed deep learning-based simple online and real-time vessel data fusion method (DeepSORVF). The DeepSORVF consists of AIS-based vessel trajectory extraction, video-based vessel trajectory extraction, and asynchronous vessel trajectory matching.

Shao et al. [32] proposed a YOLOv2-based saliency-aware network for vessel detection, which combined the salient features and coastline features to predict more accurate vessel positions. Liu et al. [14] built an enhanced YOLOv3 network to promote vessel detection in video-based maritime surveillance. To reduce the impact of poor weather environments on vessel detection, this method constructed a data enhancement strategy to improve vessel detection precision in low-light, hazy, and rainy images. Furthermore, Chen et al. [33] proposed a small vessel detection method based on an improved generative adversarial network (GAN) and a convolutional neural network (CNN). Feng et al. [4] proposed a ship detection method based on the multi-size gradient features and multi-branch support vector machine (SVM). Yang et al. [34] applied the visual object tracking and semi-supervised object segmentation to the vessel tracking task, and proposed an enhanced SiamMask network. Despite numerous deep learning-based vessel detection and tracking methods that have been proposed, these methods often suffer from detection failures or bounding box displacements when visual vessels occlude each other.

### B. AIS and Video Data Fusion

In current literature, many AIS and video data fusion methods have been proposed. For instance, Chen et al. [5] proposed a single-vessel tracking method by combining AIS and video data. In particular, this method could make the camera focus on the vessel according to the position information provided by the AIS, and use the Kalman filter to ensure the smoothness of the tracking. However, the operator fails to accurately obtain the identities and attributes of each vessel when the field of view has multiple vessels. Therefore, more researchers began to focus on the information fusion of multiple vessels. For instance, Man et al. [6] fused the AIS and video data with the Kalman filter to obtain the optimal vessel trajectory. Bloisi et al. [1] proposed an automated maritime surveillance system that replaces radar sensors with vision sensors, which can be deployed in densely populated regions. Lu et al. [7] proposed a vision and AIS fusion method, which estimated the distance and azimuth of the

detected visual vessel from the camera and fused it with the position information in the AIS data. Huang et al. [8] designed a novel multi-vessel tracking technology based on the improved single shot multi-box detector (SSD) [26] and DeepSORT [35] algorithm, and used a multi-modal data fusion algorithm to display the AIS information of visual targets. Recently, Liu et al. [9] constructed an intelligent edge-enabled shipboard navigation system based on augmented reality, deep object detection, and multi-source data fusion technologies. This system can achieve stable vessel detection under various complex weather conditions and fuse the detected vessel targets with synchronized AIS information. Although these methods use the current vessel features (such as position, azimuth, and motion) to match AIS and visual targets, the importance of historical information is ignored, which limits the accuracy of matching.

### III. DEEPSORVF: DEEP LEARNING-BASED SIMPLE ONLINE AND REAL-TIME VESSEL DATA FUSION

In this section, the details of our method will be introduced. Fig. 1 displays the flowchart of our data fusion method, including AIS-based vessel trajectory extraction, video-based vessel trajectory extraction, and asynchronous vessel trajectory matching. For the AIS data, we perform data cleaning and delayed data prediction to obtain high-quality AIS data. To guarantee that the AIS and video data are in the same coordinate system, we use the pinhole model to project the AIS data to the pixel coordinate system. For the video data, we first use the YOLOX network to detect vessel targets. To avoid the impact of vessel occlusion on the video-based trajectory, a prior knowledge-driven anti-occlusion tracking method is then used for video-based vessel trajectory extraction. During trajectory matching, we adopt the enhanced fast dynamic time warping algorithm (E-FastDTW) to calculate the similarity between trajectories and combine the Hungarian algorithm to obtain the matching results. It is worth mentioning that the matching result will be input into the video-based vessel trajectory extraction task at the next moment as prior knowledge. Based on our matching results, the AIS information, including MMSI, longitude, latitude, speed, course, heading, etc., and the



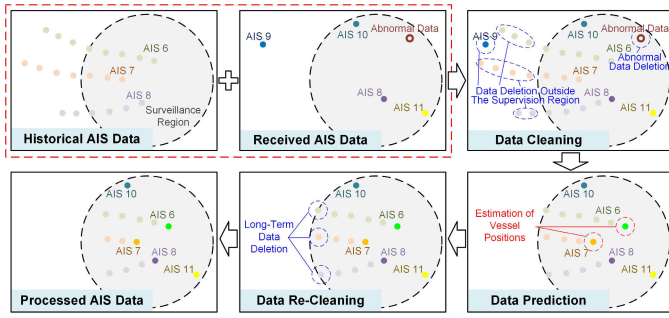


Fig. 2. The flowchart of the AIS data processing, which consists of data cleaning, data prediction, and data re-cleaning.

visual vessel can be easily fused to facilitate inland waterways surveillance.

### A. AIS-Based Vessel Trajectory Extraction

The AIS is widely used in maritime services since it can provide integrated and rich vessel information. However, due to the limitation of the AIS working principle, AIS data fails to be required in real-time. Meanwhile, some abnormal and redundant AIS information will affect the accuracy and robustness of AIS-based trajectory extraction. Furthermore, the matching of these two data becomes increasingly difficult since AIS- and video-based trajectories are, respectively, in the WGS-84 and pixel coordinate systems. Therefore, we tend to extract the AIS-based trajectory projected in the pixel coordinate system for data fusion. To achieve this goal, we first process the AIS data to generate high-quality AIS data. Subsequently, the AIS-based vessel trajectory will be obtained by the pinhole model.

1) *AIS Data Processing*: Fig. 2 displays our framework for processing AIS data. The historically processed AIS data and the AIS data received at the current moment are combined as the input. The input data is successively processed by data cleaning, data prediction, and data re-cleaning to obtain high-quality AIS data as output. The data cleaning process is used to delete the AIS data outside the supervision region and abnormal data, including the missing and abnormalities of the latitude, longitude, heading, speed, and MMSI. The data prediction module can estimate the position of vessels that have not yet received AIS information. Let  $v_{t-1}$  be the speed of the vessel at time  $T_{t-1}$ , the moving distance  $D_{\Delta t}$  at the time interval  $\Delta t = T_t - T_{t-1}$  can be expressed as  $D_{\Delta t} = v_{t-1} * \Delta t$ . According to the longitude  $\lambda_{t-1}$ , latitude  $\phi_{t-1}$ , and course  $\theta_{t-1}$  at time  $T_{t-1}$ , and the moving distance at time interval  $D_{\Delta t}$ , the longitude  $\lambda_t$  and latitude  $\phi_t$  at time  $T_t$  can be generated by the forward geodetic computations.<sup>1</sup>

2) *Vessel Positioning via Coordinate Transformation*: To fuse the high-quality AIS information and visual object, it is necessary to unify the different source data into the same coordinate system. In this work, we tend to project the AIS information in the world coordinate system (WCS) to the pixel coordinate system (PCS). Before the coordinate transformation, we first perform a Mercator projection on the

original position of the AIS information. Let  $(U, V, W)$  be the real vessel position in the 3D WCS, its 2D projection coordinate  $(x, y)$  in PCS can be obtained by

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \frac{1}{Z} \mathcal{K}_{in} \mathcal{K}_{ex} \begin{bmatrix} U \\ V \\ W \\ 1 \end{bmatrix}, \quad (1)$$

with  $Z$  being the scale factor. Here,  $\mathcal{K}_{in}$  and  $\mathcal{K}_{ex}$  are the internal and external parameter matrices of the camera, respectively. In this work, since the camera is fixed, we set the extrinsic parameter matrix  $\mathcal{K}_{ex}$  as an identity matrix. In particular, we directly use the pinhole model to estimate  $\mathcal{K}_{in}$ . Please refer to [36] for more details on the internal parameter estimation. Finally, we sequentially save the AIS data with the same MMSI into the same list in time series to build a set of all AIS-based vessel trajectories  $\mathcal{T}_{ais} = \{X_{a_1}, \dots, X_{a_i}, \dots, X_{a_I}\}$  with  $X_{a_i}$  and  $I$  being the  $i$ -th AIS trajectory and the number of AIS-based vessel trajectories.

### B. Video-Based Vessel Trajectory Extraction

Although many methods have been proposed to achieve vessel detection and tracking [4], [18], [37], it is still intractable to extract high-quality video-based vessel trajectories for data fusion. In the actual application of video-based maritime surveillance, the inevitable occlusion between vessels occurs in the cross encounter, confrontation, and chasing situations. Generally, it becomes difficult to accurately and robustly detect these vessels under the occlusion condition. Meanwhile, the corresponding appearance will be seriously affected by other vessels. To improve the quality of extracted trajectories, we propose a prior knowledge-driven anti-occlusion tracking method, as shown in Fig. 3. The AIS/video association results of the previous moment are considered to predict the bounding box of the occluded region and improve the visual trajectory quality of the occluded vessel.

Specifically, we first adopt the YOLOX network [30] to detect the visual vessel object and get a set of bounding boxes, which can be given by

$$\text{Boxes} = \{box_1, \dots, box_l, \dots, box_L\}, \quad (2)$$

where  $L$  denotes the number of bounding boxes.  $box_l = [x_{tl}, y_{tl}, x_{br}, y_{br}]^T$  is the location of the  $l$ -th bounding box, with  $(x_{tl}, y_{tl})$  and  $(x_{br}, y_{br})$  being the pixel indexes of the top-left and bottom-right points.

1) *Bounding Box Removal in the Occlusion Areas*: Before the tracking, the results of the previous moment are input as the prior knowledge. Firstly, a set of the occlusion areas OAR will be used, which depends on the ratio of the occlusion area to the bounding boxes. The judgment metric of the occlusion area can be expressed as follows

$$\frac{S_o}{\min(S_1, \dots, S_r, \dots, S_R)} > \omega, \quad (3)$$

where  $S_o$  is the area of the occluded part,  $S_r$  is the area of the  $r$ -th occluded bounding box,  $R$  is the number of occluded bounding boxes,  $\omega$  represents the anti-occlusion threshold. When the ratio of  $S_o$  to the minimum bounding box of the

<sup>1</sup>We exploit the `pyproj.Geod.fwd` function to implement the forward geodetic computations.

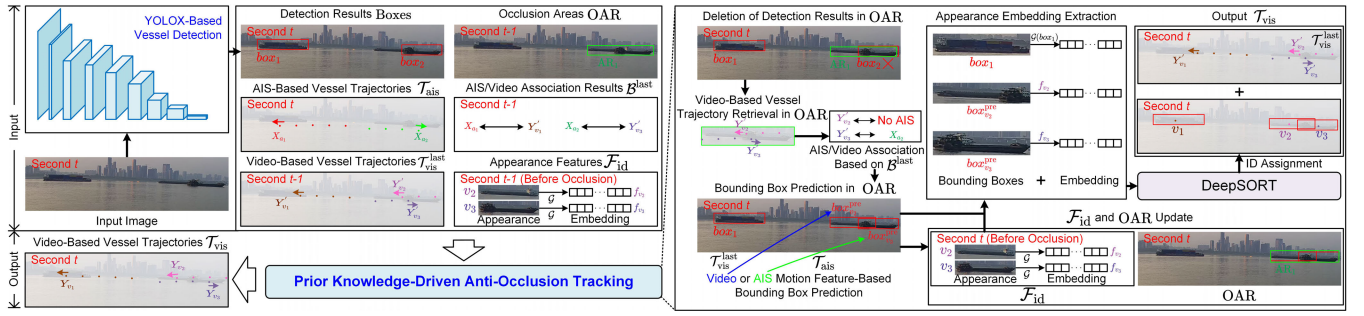


Fig. 3. The flowchart of anti-occlusion tracking method for video-based vessel trajectory extraction. Note that  $\mathcal{G}$  is the wide residual network-based appearance feature extractor. The extraction of AIS-based vessel trajectories  $\mathcal{T}_{\text{ais}}$  has been introduced in Section III-A. The generation of Boxes, OAR,  $\mathcal{T}_{\text{vis}}^{\text{last}}$ , and  $\mathcal{F}_{\text{id}}$  will be mentioned in Section III-B. The generation of AIS/video association results  $\mathcal{B}^{\text{last}}$  will be described in Section III-C.

occlusion region exceeds  $\omega$ , we will store the location of the smallest rectangle box AR which can contain all occluded bounding boxes into the OAR. Meanwhile, the AIS-based vessel trajectories at the current moment  $\mathcal{T}_{\text{ais}}$  and the video-based vessel trajectories at the previous moment  $\mathcal{T}_{\text{vis}}^{\text{last}}$  are also used as prior information, which can be given by

$$\begin{cases} \mathcal{T}_{\text{ais}} = \{X_{a_1}, \dots, X_{a_i}, \dots, X_{a_I}\}, \\ \mathcal{T}_{\text{vis}}^{\text{last}} = \{Y'_{v_1}, \dots, Y'_{v_j}, \dots, Y'_{v_J}\}, \end{cases} \quad (4)$$

where  $X_{a_i}$  and  $Y'_{v_j}$  represent the trajectory series of the  $i$ -th AIS target and the  $j$ -th visual target, respectively,  $I$  and  $J$  are the numbers of AIS- and video-based vessel trajectories, respectively. Besides, we consider the AIS/video association results  $\mathcal{B}^{\text{last}}$  at the previous moment and the vessel appearance embedding  $\mathcal{F}_{\text{id}}$  before the occlusion, which can be given by

$$\begin{cases} \mathcal{B}^{\text{last}} = \{\dots, [a_i, v_j], \dots\}, \\ \mathcal{F}_{\text{id}} = \{\dots, f_{v_b}, \dots\}, \end{cases} \quad (5)$$

where  $[a_i, v_j]$  means that the  $i$ -th AIS target  $a_i$  and  $j$ -th visual target  $v_j$  are successfully associated,  $f_{v_b}$  is the appearance embedding of the  $b$ -th visual target before occlusion.<sup>2</sup> We will detailedly describe the generation of AIS/video association results  $\mathcal{B}^{\text{last}}$  in Section III-C.

2) *Occluded Bounding Box Prediction*: For the anti-occlusion tracking, the detection results located in the occlusion area OAR are removed to avoid the mis-detection caused by the vessel overlapping. Based on the fusion results at the previous moment, the corresponding AIS information of the occluded visual vessel is available. Therefore, the location of the occluded vessel's bounding box at the current moment  $\text{box}^{\text{pre}}$  can be estimated by

$$\text{box}^{\text{pre}} = \begin{bmatrix} x_{\text{tl}}^{\text{pre}} \\ y_{\text{tl}}^{\text{pre}} \\ x_{\text{br}}^{\text{pre}} \\ y_{\text{br}}^{\text{pre}} \end{bmatrix} = \begin{bmatrix} \Delta x_{\text{ais}} \\ \Delta y_{\text{ais}} \\ \Delta x_{\text{ais}} \\ \Delta y_{\text{ais}} \end{bmatrix} + \begin{bmatrix} x_{\text{tl}}^{\text{last}} \\ y_{\text{tl}}^{\text{last}} \\ x_{\text{br}}^{\text{last}} \\ y_{\text{br}}^{\text{last}} \end{bmatrix}, \quad (6)$$

where  $(x_{\text{tl}}^{\text{last}}, y_{\text{tl}}^{\text{last}})$  and  $(x_{\text{br}}^{\text{last}}, y_{\text{br}}^{\text{last}})$  are the pixel indexes of the top-left and bottom-right points of the previous bounding box, respectively.  $\Delta x_{\text{ais}}$  and  $\Delta y_{\text{ais}}$  are the horizontal and vertical motion speeds, respectively, which are equal to the

horizontal and vertical differences between the AIS pixel positions generated by Eq. (1) at the current and previous moments.

For the occluded vessels without the corresponding AIS information, the bounding box will be predicted via the visual motion features. The prediction result can also be given by variants based on Eq. (6). The horizontal and vertical motion speeds  $(\Delta x_{\text{ais}}, \Delta y_{\text{ais}})$  are replaced with the visual trajectory-based horizontal and vertical motion speeds  $(\Delta x_{\text{vis}}, \Delta y_{\text{vis}})$ , which can be calculated by

$$\begin{cases} \Delta x_{\text{vis}} = \frac{x_{t-1} - x_{t-\delta}}{\delta}, \\ \Delta y_{\text{vis}} = \frac{y_{t-1} - y_{t-\delta}}{\delta}, \end{cases} \quad (7)$$

where  $(x_{t-1}, y_{t-1})$  and  $(x_{t-\delta}, y_{t-\delta})$  denote the points of the video-based vessel trajectory at the previous moment and the previous  $\delta$  moment, respectively.

After prediction, we will then update the OAR and  $\mathcal{F}_{\text{id}}$ . Based on the predicted detection box position, the occlusion area list OAR will be updated via Eq. (3) for the anti-occlusion tracking at the next moment. For the update of  $\mathcal{F}_{\text{id}}$ , we first set up the occluded visual target as  $v_j$ . If  $f_{v_j}$  exists in the original  $\mathcal{F}_{\text{id}}$ , we directly store  $f_{v_j}$  into the new  $\mathcal{F}_{\text{id}}$ ; otherwise, the appearance embedding of  $v_j$  at the previous moment will be stored in the new  $\mathcal{F}_{\text{id}}$ . Then, we employ a wide residual network  $\mathcal{G}$  to extract the vessel appearance embedding in normal bounding boxes, and assign the vessel appearance embedding before occlusion in the  $\mathcal{F}_{\text{id}}$  to occluded bounding boxes. Finally, the bounding boxes and the corresponding vessel appearance embedding are jointly input into the DeepSORT [35] for generating the video-based vessel trajectories at the current moment  $\mathcal{T}_{\text{vis}}$ .

3) *DeepSORT-Based Vessel Tracking*: It is worth mentioning that DeepSORT employs two metrics to resolve ID assignment issues. Firstly, the Mahalanobis distances between the predicted Kalman states and the newly arrived locations are calculated as the location similarity metrics. Moreover, the cosine distances between the appearance embedding are calculated as the appearance similarity metrics. In our method, the appearance features of the occluded vessels are kept consistent with the latest extractions before the occlusion. Therefore, as long as the predicted bounding box is close to the prediction of Kalman filters, the ID of occluded

<sup>2</sup>The vessel appearance embedding is extracted by the wide residual network [38] in the DeepSORT.

**Algorithm 1** Anti-Occlusion Vessel Tracking

---

**Input:** OAR: A set of all occlusion areas; Boxes: A set of all bounding boxes detected by YOLOX network;  $\mathcal{T}_{\text{ais}}$ : A set of all AIS-based vessel trajectories at the current moment;  $\mathcal{T}_{\text{vis}}^{\text{last}}$ : A set of all video-based vessel trajectories at the previous moment;  $\mathcal{B}^{\text{last}}$ : A set of all AIS/video association results at the previous moment;  $\mathcal{F}_{\text{id}}$ : A set of all occluded vessel appearance features before the occlusion;

**Output:**  $\mathcal{T}_{\text{vis}}$ : A set of all video-based vessel trajectories at the current moment;

- 1 **Initialization:** Boxes<sup>pre</sup>: A empty set to save all predict bounding boxes of the occluded vessels;  $\mathcal{G}$ : A wide residual network-based appearance feature extractor;  $\mathcal{I}$ : A empty set to save all DeepSORT input data;

```
// Step 1. Bounding box removal in the
// occlusion areas.
2 for boxl in Boxes do
3   for AR in OAR do
4     if boxl locates in AR then
5       Remove the boxl from Boxes;
6       break;
// Step 2. Occluded bounding box
// prediction.
7 for Y'vj in  $\mathcal{T}_{\text{vis}}^{\text{last}}$  do
8   for AR in OAR do
9     if The center point of the bounding box at the
10      previous moment in Y'vj locates in AR then
11       Search the matched [·, vj] from  $\mathcal{B}^{\text{last}}$ ;
12       if exist [ai, vj] then
13         Predict the bounding box boxvjpre by
14         AIS-based vessel trajectory Xai in  $\mathcal{T}_{\text{ais}}$ ;
15       else
16         Predict the bounding box boxvjpre by
17         video-based vessel trajectory Y'vj;
18       Add the boxvjpre to Boxespre;
19       break;
17 Update  $\mathcal{F}_{\text{id}}$  and OAR;
// Step 3. DeepSORT-based vessel
// tracking.
18 for boxl in Boxes do
19   Add [boxl,  $\mathcal{G}(\text{box}_l)$ ] to  $\mathcal{I}$ ;
20 for boxvjpre in Boxespre do
21   for fvb in  $\mathcal{F}_{\text{id}}$  do
22     if vj = vb then
23       Add [boxvjpre, fvb] to  $\mathcal{I}$ ;
24       break;
25 Run DeepSORT with  $\mathcal{I}$ ;
26 Add the results of DeepSORT to  $\mathcal{T}_{\text{vis}}^{\text{last}}$  for generating the
// video-based vessel trajectory at the current moment  $\mathcal{T}_{\text{vis}}$ ;
```

---

vessels will not be assigned incorrectly. The pseudo-code of the proposed anti-occlusion tracking method is shown in Algorithm 1.

*C. Asynchronous Vessel Trajectory Matching*

In this section, we propose a simple yet effective trajectory matching method to fuse the AIS- and video-based asynchronous vessel trajectories. Firstly, we adopt an enhanced fast dynamic time warping (E-FastDTW) algorithm considering the direction to calculate the similarity of AIS- and video-based vessel trajectories. Based on the similarity measure result, the Hungarian algorithm is employed to generate the optimal matching result. To improve the stability and robustness of data fusion and reduce the computational cost, we employ an AIS/video association mechanism. When the number of successful pairings of two trajectories exceeds a pre-determined threshold, the AIS- and video-based vessel trajectories will be associated directly without similarity evaluation.

1) *Trajectory Similarity Measure via E-FastDTW:* For trajectory-based data fusion, it is an important prerequisite to determine the similarities between the AIS- and video-based vessel trajectories. The Euclidean distance is a simple but effective similarity calculation method. However, it requires that the two trajectories to be matched have the same length. Meanwhile, the Euclidean distance considers that two similar trajectories with only a slight shift in the time axis are significantly different. Therefore, dynamic time warping (DTW) has been proposed for ignoring this shift [39]. Suppose we have two trajectories  $X$  and  $Y$  of length  $P$  and  $Q$  respectively, represented as

$$\begin{cases} X = m_1, m_2, \dots, m_p, \dots, m_P, \\ Y = n_1, n_2, \dots, n_q, \dots, n_Q. \end{cases} \quad (8)$$

Based on the two trajectories, the DTW constructs a  $P \times Q$  alignment matrix  $d$  where  $d(p, q)$  is the Euclidean distance between the points  $m_p$  and  $n_q$ . Then, a warp path  $W$  is defined to construct the mapping between  $X$  and  $Y$ , which can be written by

$$W = w_1, w_2, \dots, w_c, \dots, w_C, \quad (9)$$

with  $C$  being the length of  $W$ , and  $\max\{P, Q\} \leq C < P + Q$ . In particular, the warp path  $W$  has three restrictions. For the sake of better understanding, we define the  $(c-1)$ -th and the  $c$ -th elements of  $W$  as  $w_{c-1} = (p', q')$  and  $w_c = (p, q)$ . These three constraints for warp path can be defined as follows:

- *Restriction 1:* The 1-st and the  $C$ -th elements of  $W$  are  $w_1 = (1, 1)$  and  $w_C = (P, Q)$ , respectively.
- *Restriction 2:* The adjacent elements of the warp path  $W$  can only contain the adjacent coordinate points, including the diagonal adjacent. Therefore, the  $w_{c-1}$  can only be one of  $\{(p-1, q), (p, q-1), (p-1, q-1)\}$ .
- *Restriction 3:* The elements of the warp path  $W$  are monotonically increasing in time, i.e.,  $p' \leq p$  and  $q' \leq q$ .

Under the premise of satisfying the above three constraints, DTW only focuses on the path with the minimum cumulative distance of alignment matrix elements corresponding to all points [39]. Meanwhile, the included angle  $\varphi$  between the starting and ending points of  $X$  and  $Y$  is also considered. Finally, the similarity value  $S(X, Y)$  between  $X$  and  $Y$  calculated by our proposed E-FastDTW can be written



---

**Algorithm 2** Asynchronous Vessel Trajectory Matching
 

---

**Input:**  $\mathcal{T}_{\text{ais}}$ : A set of all AIS-based vessel trajectories;  $\mathcal{T}_{\text{vis}}$ : A set of all video-based vessel trajectories;  $\mathcal{M}^{\text{last}}$ : A set of all AIS/video numbers of matches at the previous moment;  $\mathcal{B}^{\text{last}}$ : A set of all AIS/video association result at the previous moment;

**Output:**  $\mathcal{M}$ : A set of all AIS/video numbers of matches at the current moment;  $\mathcal{B}$ : A set of all AIS/video association result at the current moment;

1 **Initialization:**  $d(i, j)$ : The Euclidean distance between the last trajectory points of  $X_{a_i}$  and  $Y_{v_j}$ ;  $M_s$ : An empty trajectory similarity matrix;  $\mathcal{O}_{\text{res}}$ : An empty set to save the matching results;  $D_{\text{max}}$ : The maximum matching distance;  $Mat_{\text{min}}$ : The minimum number of matches;  $T_{\text{max}}$ : The maximum time threshold;  $S$ : The E-FastDTW trajectory similarity measurement operator;

// Step 1. Trajectory similarity measure.

2 **for**  $X_{a_i}$  **in**  $\mathcal{T}_{\text{ais}}$  **do**

3     **for**  $Y_{v_j}$  **in**  $\mathcal{T}_{\text{vis}}$  **do**

4         **if**  $d(i, j) > D_{\text{max}}$  **then**

5              $M_s(i, j) = +\infty$ ;

6         **else if**  $[a_i, :]$  **or**  $[:, v_j]$  **in**  $\mathcal{B}^{\text{last}}$  **then**

7             **if**  $[a_i, v_j]$  **in**  $\mathcal{B}^{\text{last}}$  **then**

8                  $M_s(i, j) = -\infty$ ;

9             **else**

10                  $M_s(i, j) = +\infty$ ;

11         **else**

12              $M_s(i, j) = S(X_{a_i}, Y_{v_j})$  by Eq. (10);

// Step 2. Matching result generation.

13 Using the Hungarian algorithm to calculate  $M_s$  for obtaining the matching result  $\mathcal{O}_{\text{res}} = \{\dots, [a_i, v_j], \dots\}$ ;

14 **for**  $[a_i, v_j]$  **in**  $\mathcal{O}_{\text{res}}$  **do**

15     **if**  $z_{a_i, v_j}$  **in**  $\mathcal{M}^{\text{last}}$  **then**

16         Add  $z_{a_i, v_j} = z_{a_i, v_j} + 1$  to  $\mathcal{M}$ ;

17     **else**

18         Add  $z_{a_i, v_j} = 1$  to  $\mathcal{M}$ ;

19 **for**  $z_{a_i, v_j}$  **in**  $\mathcal{M}^{\text{last}}$  **do**

20     **if**  $[a_i, v_j]$  **not in**  $\mathcal{O}_{\text{res}}$  **and the time interval between the last matching moment of**  $[a_i, v_j]$  **and the current moment**  $< T_{\text{max}}$  **then**

21         Add  $z_{a_i, v_j}$  to  $\mathcal{M}$ ;

// Step 3. Association result generation.

22 **for**  $z_{a_i, v_j}$  **in**  $\mathcal{M}$  **do**

23     **if**  $z_{a_i, v_j} > Mat_{\text{min}}$  **then**

24         Add  $[a_i, v_j]$  to  $\mathcal{B}$ ;

---

as follows

$$S(X, Y) = Dis(W) \cdot e^\varphi = \min\left\{\sum_{c=1}^C d(w_{cp}, w_{cq})\right\} \cdot e^\varphi, \quad (10)$$

where  $d(w_{cp}, w_{cq})$  is the Euclidean distance between two data points corresponding to the  $c$ -th element in the warp path

$W$ ,  $Dis(W)$  denotes the sum of all  $d(w_{cp}, w_{cq})$  in the warp path  $W$ . To find the desired unique warp path, the DTW adopts the dynamic programming strategy. The cumulative distance  $\mathcal{D}(p, q)$  between  $m_p$  and  $n_q$  is the sum of the minimum cumulative distance of three previous possible warp path elements and the Euclidean distance  $d(p, q)$  between the points  $m_p$  and  $n_q$ , which can be mathematically written as

$$\begin{aligned} \mathcal{D}(p, q) &= d(p, q) + \min\{\mathcal{D}(p-1, q), \\ &\quad \mathcal{D}(p, q-1), \mathcal{D}(p-1, q-1)\}. \end{aligned} \quad (11)$$

Furthermore, we also adopt the multi-level approach used in the FastDTW to speed up the time series similarity search and reduce the computational complexity. Please refer to [40] for more details on the multi-level approach.

2) *Trajectory Matching:* In this work, we propose a novel matching method with higher precision and less computation. In particular, we will match and associate the AIS-based vessel trajectories  $\mathcal{T}_{\text{ais}}$  mentioned in Section III-A, and the video-based vessel trajectories  $\mathcal{T}_{\text{vis}}$  mentioned in Section III-B, which can be defined as follows

$$\begin{cases} \mathcal{T}_{\text{ais}} = \{X_{a_1}, \dots, X_{a_i}, \dots, X_{a_I}\}, \\ \mathcal{T}_{\text{vis}} = \{Y_{v_1}, \dots, Y_{v_j}, \dots, Y_{v_J}\}, \end{cases} \quad (12)$$

where  $X_{a_i}$  and  $Y_{v_j}$  represent the trajectories of the  $i$ -th AIS target  $a_i$  and the  $j$ -th visual target  $v_j$ , respectively,  $I$  and  $J$  are the numbers of AIS- and video-based vessel trajectories, respectively. Furthermore, the numbers of AIS/video matches  $\mathcal{M}^{\text{last}}$  and association results  $\mathcal{B}^{\text{last}}$  at the previous moment are also considered as input, i.e.,

$$\begin{cases} \mathcal{M}^{\text{last}} = \{\dots, z_{a_i, v_j}, \dots\}, \\ \mathcal{B}^{\text{last}} = \{\dots, [a_i, v_j], \dots\}, \end{cases} \quad (13)$$

where  $z_{a_i, v_j}$  is the number of successful matches of  $X_{a_i}$  and  $Y_{v_j}$ ,  $[a_i, v_j]$  means that  $a_i$  and  $v_j$  have been associated together. In the similarity measure, it is obviously time-consuming and intractable to adopt the E-FastDTW for calculating the similarity between all trajectories at each moment. Inspired by the DeepSORT algorithm, we propose a trajectory association mechanism to solve these issues. In particular, if two trajectories have been recorded in the  $\mathcal{B}^{\text{last}}$ , the two trajectories are directly matched by default without similarity measurement with other trajectories. Subsequently, we perform the similarity measure between all trajectories and construct a similarity matrix  $M_s$  of size  $I \times J$ , where  $M_s(i, j)$  represents the similarity value of  $X_{a_i}$  and  $Y_{v_j}$ . In particular, when the Euclidean distance between the last trajectory points of  $X_{a_i}$  and  $Y_{v_j}$  exceeds the maximum matching distance  $D_{\text{max}}$ , we consider the two trajectories to be completely different and set  $M_s(i, j) = +\infty$ . When the binding trajectory pair  $[a_i, v_j]$  exists in the  $\mathcal{B}^{\text{last}}$ , we set  $M_s(i, j) = -\infty$  and set the values of other horizontal and vertical positions to positive infinity. For other ordinary trajectory pairs that do not satisfy the above conditions, we employ Eq. (10) (i.e., E-FastDTW) to calculate the trajectory similarity. After obtaining the similarity matrix  $M_s$ , we adopt the Hungarian optimization algorithm [41] to find the optimal matching result  $\mathcal{O}_{\text{res}}$ , which contains the

matching trajectory pair information, i.e.,

$$\mathcal{O}_{\text{res}} = \{\dots, [a_i, v_j], \dots\}, \quad (14)$$

where  $[a_i, v_j]$  means that  $a_i$  and  $v_j$  are matched together. Then, we will generate the AIS/video matching results  $\mathcal{M}$  and association results  $\mathcal{B}$  at the current moment. More specifically, we iterate through all matching trajectory pairs in the  $\mathcal{O}_{\text{res}}$ . If the number of matching times  $z_{a_i, v_j}$  of trajectory pair  $[a_i, v_j]$  in the  $\mathcal{O}_{\text{res}}$  exists in the  $\mathcal{M}^{\text{last}}$ , we will store  $(z_{a_i, v_j} + 1)$  to  $\mathcal{M}$ ; otherwise, 1 will be stored to  $\mathcal{M}$ . In addition, we save the number of matching times  $z_{a_i, v_j}$  for some trajectory pairs directly from  $\mathcal{M}^{\text{last}}$  into  $\mathcal{M}$ . These  $z_{a_i, v_j}$  need to satisfy two conditions, which can be defined as follows:

- $z_{a_i, v_j}$  must exist in  $\mathcal{M}^{\text{last}}$  but  $[a_i, v_j]$  is not in  $\mathcal{O}_{\text{res}}$ .
- The time interval between the last matching moment and the current moment is less than  $T_{\text{max}}$ .

For the generation of the AIS/video association result, we set a minimum number of matches  $Mat_{\text{min}}$  as a threshold to ensure that the association information is accurate. When  $z_{a_i, v_j}$  in the  $\mathcal{M}$  is greater than  $Mat_{\text{min}}$ , we will store  $[a_i, v_j]$  into  $\mathcal{B}$ . The pseudo-code of the proposed trajectory matching method is shown in Algorithm 2.

#### D. Implementation Details

This section mainly introduces the detailed settings of the proposed data fusion method. In particular, our method is implemented on the Python 3.7 platform. All experiments and tests are conducted on a PC with Intel Core i5-10600KF CPU @ 4.10GHz and Nvidia RTX A4000 GPU. To meet the requirement of real-time processing while ensuring accurate fusion, our method only executes one processing per second. For the AIS-based vessel trajectory extraction, we delete the data more than two nautical miles from the camera and set the maximum storage time to two minutes. For the vessel detection task, we collect 20k images containing vessel objects as the training dataset. In training, we set the epoch to 100 and employ the Adam algorithm as the optimizer. The initial learning rates for the first 50 and last 50 epochs are  $10^{-3}$  and  $10^{-4}$ , respectively. For the video-based vessel trajectory extraction, we set the occlusion area threshold  $\omega = 0$  and the time span of visual motion feature extraction  $\delta = 5s$ . For the AIS and video data fusion, we set the maximum matching distance  $D_{\text{max}}$  as the half of the horizontal size of the image, the minimum number of matching times  $Mat_{\text{min}} = 15$ , and the maximum time threshold  $T_{\text{max}} = 15s$ .

### IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we conduct massive experiments on vessel detection, vessel tracking, and data fusion to quantitatively evaluate the performance of our proposed method. The running time analysis is also carried out to verify its practicality.

#### A. Benchmark Dataset

In this section, we construct a benchmark dataset for vessel detection, tracking, and data fusion (named FVessel) containing 26 videos and the corresponding AIS data captured by the

TABLE I

DETAILS OF THE FVESSEL DATASET. THE ‘‘TOO’’, ‘‘NOV’’, AND ‘‘NOA’’ ARE THE TIMES OF OCCLUSIONS, THE TOTAL NUMBER OF VESSELS, AND THE NUMBER OF VESSELS WITH AIS, RESPECTIVELY

Video	Video Length	Type	Weather	TOO	NOV	NOA
video-01	10m07s	Bridge	Low-light	2	5	4
video-02	19m52s	Bridge	Sunny	6	7	6
video-03	19m14s	Riverside	Sunny	6	5	5
video-04	06m10s	Riverside	Sunny	0	1	1
video-05	15m01s	Riverside	Sunny	2	5	5
video-06	12m49s	Riverside	Sunny	2	4	4
video-07	03m38s	Riverside	Sunny	1	2	2
video-08	16m05s	Riverside	Sunny	3	6	5
video-09	05m25s	Riverside	Sunny	0	1	1
video-10	11m17s	Bridge	Cloudy	2	3	1
video-11	05m18s	Riverside	Sunny	1	3	3
video-12	07m19s	Riverside	Sunny	1	4	4
video-13	12m58s	Riverside	Sunny	5	6	6
video-14	03m58s	Riverside	Sunny	3	4	4
video-15	10m46s	Riverside	Sunny	0	4	4
video-16	05m05s	Riverside	Sunny	0	1	1
video-17	08m08s	Riverside	Sunny	1	2	2
video-18	23m57s	Riverside	Sunny	10	10	6
video-19	11m28s	Riverside	Low-light	0	2	2
video-20	14m10s	Riverside	Low-light	0	3	3
video-21	24m01s	Riverside	Low-light	4	7	6
video-22	02m40s	Riverside	Low-light	0	2	1
video-23	19m24s	Riverside	Sunny	2	4	4
video-24	08m39s	Riverside	Sunny	2	3	3
video-25	24m05s	Riverside	Sunny	4	8	8
video-26	07m26s	Riverside	Sunny	0	5	5



Fig. 4. Some samples of the FVessel dataset. This dataset contains massive images and videos captured on the bridge region and riverside under sunny, cloudy, and low-light conditions.

HIKVISION DS-2DC4423IW-D dome camera and Saiyang AIS9000-08 Class-B AIS receiver on the Wuhan Segment of the Yangtze River.<sup>3</sup> As shown in Fig. 4, these videos were captured under many locations (e.g., bridge region and riverside) and various weather conditions (e.g., sunny, cloudy, and low-light). Table I displays more details about the FVessel dataset, including the video length, collection location, weather condition, the times of occlusions, the total number of vessels, and the number of vessels with AIS. To verify the superiority of the proposed module, we intercept ten clips existing the vessel occlusion from the FVessel dataset for comparison experiments on vessel detection, tracking, and data fusion. More detailed information on the test dataset can be found in Table II.

#### B. Experiments on Data Fusion

In this section, we implement the data fusion experiment to compare various methods, i.e., Euclidean distance-based data fusion (EDDF), multi-source data fusion (MSDF) [9], multi-modal data-based ship tracking (MMDST) [8],

<sup>3</sup>The vessel is uniquely identified by a particular MMSI in raw AIS data. To protect the privacy, the MMSI for each vessel has been replaced with a random number in our dataset.



TABLE II

DETAILS OF THE DATASET USED IN THE VESSEL DETECTION, VESSEL TRACKING, AND DATA FUSION EXPERIMENTS. THE “TOO”, “NOV”, AND “NOA” ARE THE TIMES OF OCCLUSIONS, THE TOTAL NUMBER OF VESSELS, AND THE NUMBER OF VESSELS WITH AIS, RESPECTIVELY

Video	Video Length	Type	Weather	TOO	NOV	NOA
clip-01	01m51s	Bridge	Cloudy	1	2	2
clip-02	01m36s	Riverside	Sunny	1	2	2
clip-03	03m42s	Riverside	Low-light	1	2	2
clip-04	02m42s	Riverside	Low-light	2	3	3
clip-05	03m05s	Riverside	Sunny	2	3	3
clip-06	02m38s	Riverside	Sunny	2	3	3
clip-07	11m10s	Riverside	Sunny	2	4	4
clip-08	05m07s	Riverside	Sunny	1	2	2
clip-09	08m39s	Riverside	Sunny	2	3	3
clip-10	02m46s	Riverside	Sunny	2	4	4

DeepSORVF (w/o) without the anti-occlusion strategy, and DeepSORVF. In particular, the EDDF calculates the Euclidean distance between the pixel position points at the current moment for similarity measurement and employs the near-matching mechanism. For the point matching-based MSDF and MMDST, we only replace our asynchronous vessel trajectory matching part with its corresponding matching method to compare the fusion effect under the premise of consistent detectors. Furthermore, all methods only process data once per second.

1) *Evaluation Metric*: To evaluate the performance of data fusion, we first use a variant of multi-object tracking accuracy (MOTA) [42] as the evaluation metric and name it multi-object fusion accuracy (MOFA), i.e.,

$$MOFA = 1 - \frac{FN_{mmsi} + FP_{mmsi}}{GT_{mmsi}}, \quad (15)$$

where  $mmsi$  represents the identity of the vessel of interest (MMSI),  $FP_{mmsi}$ ,  $FN_{mmsi}$ , and  $GT_{mmsi}$  are the number of the MMSI false positive, MMSI false negative, and MMSI ground truth, respectively. Furthermore, the identification precision (IDP), identification recall (IDR), and identification F1 ( $IDF_1$ ) are also employed as evaluation metrics. The IDP, IDR, and  $IDF_1$  can be given by

$$IDP = \frac{TP_{id}}{TP_{id} + FP_{id}}, \quad (16)$$

$$IDR = \frac{TP_{id}}{TP_{id} + FN_{id}}, \quad (17)$$

$$IDF_1 = \frac{2TP_{id}}{2TP_{id} + FP_{id} + FN_{id}}, \quad (18)$$

where  $TP_{id}$ ,  $FP_{id}$ , and  $FN_{id}$  are the numbers of the ID true positive, ID false positive, and ID false negative, respectively. In particular, the  $id$  is replaced with the identity of the vessel of interest (MMSI) in the data fusion evaluation. Please refer to [42] and [43] for more details on the MOTA, IDP, IDR, and  $IDF_1$ . Generally, higher MOFA, IDP, IDR, and  $IDF_1$  mean better fusion performance.

2) *Fusion Results on Ten Clips*: Table III displays the evaluation results on all clips. It can be found that EDDF and MSDF perform poorly. Especially for clip-04, the MOFA is only 54.82% for EDDF and 53.73% for MSDF. The poor fusion effect stems from the fact that these methods only consider the current information without associating the historical feature.

TABLE III

MOFA, IDP, IDR, AND  $IDF_1$  RESULTS OF DATA FUSION FOR THE TEN CLIPS FROM TABLE II. (UNIT: %)

Video	Methods	MOFA ↑	IDP ↑	IDR ↑	$IDF_1$ ↑
clip-01	EDDF	69.27	88.67	82.57	84.31
	MSDF [9]	68.35	88.61	82.11	83.84
	MMDST [8]	94.95	<b>100.00</b>	95.41	97.42
	DeepSORVF (w/o)	95.41	<b>100.00</b>	95.41	97.65
	DeepSORVF	<b>99.54</b>	<b>100.00</b>	<b>99.54</b>	<b>99.77</b>
clip-02	EDDF	69.68	89.41	80.85	84.68
	MSDF [9]	69.68	89.41	80.85	84.68
	MMDST [8]	84.04	<b>100.00</b>	87.77	92.18
	DeepSORVF (w/o)	87.77	<b>100.00</b>	87.77	93.48
	DeepSORVF	<b>99.47</b>	<b>100.00</b>	<b>99.47</b>	<b>99.73</b>
clip-03	EDDF	79.29	96.09	87.28	89.39
	MSDF [9]	81.66	96.09	87.28	90.49
	MMDST [8]	92.60	<b>100.00</b>	92.60	96.16
	DeepSORVF (w/o)	90.53	<b>100.00</b>	90.53	95.03
	DeepSORVF	<b>94.67</b>	<b>100.00</b>	<b>94.67</b>	<b>97.26</b>
clip-04	EDDF	54.82	90.93	70.39	75.89
	MSDF [9]	53.73	90.66	72.37	75.95
	MMDST [8]	82.24	98.99	86.18	90.66
	DeepSORVF (w/o)	85.53	99.00	86.40	92.27
	DeepSORVF	<b>97.59</b>	<b>99.33</b>	<b>98.25</b>	<b>98.79</b>
clip-05	EDDF	60.34	91.71	80.17	81.39
	MSDF [9]	57.78	91.13	78.89	80.09
	MMDST [8]	79.74	<b>99.53</b>	89.55	91.01
	DeepSORVF (w/o)	86.78	98.34	88.27	93.03
	DeepSORVF	<b>92.96</b>	98.44	<b>94.46</b>	<b>96.41</b>
clip-06	EDDF	69.19	98.67	82.91	85.06
	MSDF [9]	71.15	98.66	82.63	85.88
	MMDST [8]	82.35	<b>99.37</b>	87.96	91.41
	DeepSORVF (w/o)	83.47	98.69	84.59	91.10
	DeepSORVF	<b>92.16</b>	98.81	<b>93.28</b>	<b>95.97</b>
clip-07	EDDF	76.92	95.44	88.51	89.08
	MSDF [9]	75.74	95.21	87.92	88.48
	MMDST [8]	85.85	<b>98.74</b>	92.73	93.37
	DeepSORVF (w/o)	90.18	98.42	91.65	94.91
	DeepSORVF	<b>93.03</b>	98.46	<b>94.50</b>	<b>96.44</b>
clip-08	EDDF	86.63	95.94	93.41	94.21
	MSDF [9]	87.38	95.57	93.41	94.21
	MMDST [8]	96.61	<b>100.00</b>	97.55	98.48
	DeepSORVF (w/o)	97.18	<b>100.00</b>	97.18	98.57
	DeepSORVF	<b>98.49</b>	<b>100.00</b>	<b>98.49</b>	<b>99.24</b>
clip-09	EDDF	79.08	96.42	87.91	89.61
	MSDF [9]	81.80	96.42	87.77	89.47
	MMDST [8]	92.66	<b>99.15</b>	94.57	96.53
	DeepSORVF (w/o)	92.93	98.86	94.02	96.38
	DeepSORVF	<b>94.70</b>	98.34	<b>96.33</b>	<b>97.32</b>
clip-10	EDDF	87.56	98.05	89.33	93.49
	MSDF [9]	88.44	98.54	89.78	93.95
	MMDST [8]	88.44	<b>100.00</b>	89.56	93.94
	DeepSORVF (w/o)	89.56	<b>100.00</b>	89.56	94.49
	DeepSORVF	<b>97.78</b>	<b>100.00</b>	<b>97.78</b>	<b>98.88</b>
Average	EDDF	73.28	94.13	84.33	86.71
	MSDF [9]	73.57	94.03	84.30	86.70
	MMDST [8]	87.95	<b>99.60</b>	91.39	94.12
	DeepSORVF (w/o)	89.93	99.33	90.54	94.69
	DeepSORVF	<b>96.04</b>	99.34	<b>96.68</b>	<b>97.98</b>

By considering the displacement direction of AIS- and video-based vessel trajectories, the MMDST greatly improves the fusion effect. However, the two DeepSORVFs based on the vessel motion trajectory matching have better performance by comparison. Particularly after implementing the anti-occlusion strategy, the performance of our DeepSORVF has improved considerably across all metrics.

To provide a more understandable explanation, we display two examples of data fusion obtained by MSDF, MMDST, DeepSORVF (w/o), and DeepSORVF shown in Figs. 5 and 6. Specifically, Fig. 5 displays the visualized data fusion result captured by the bridge region camera. Since the MSDF only considers the vessel characteristic at the current moment, the vessel information is more likely to be matched incorrectly. In the 80th second, the MSDF, MMDST, and

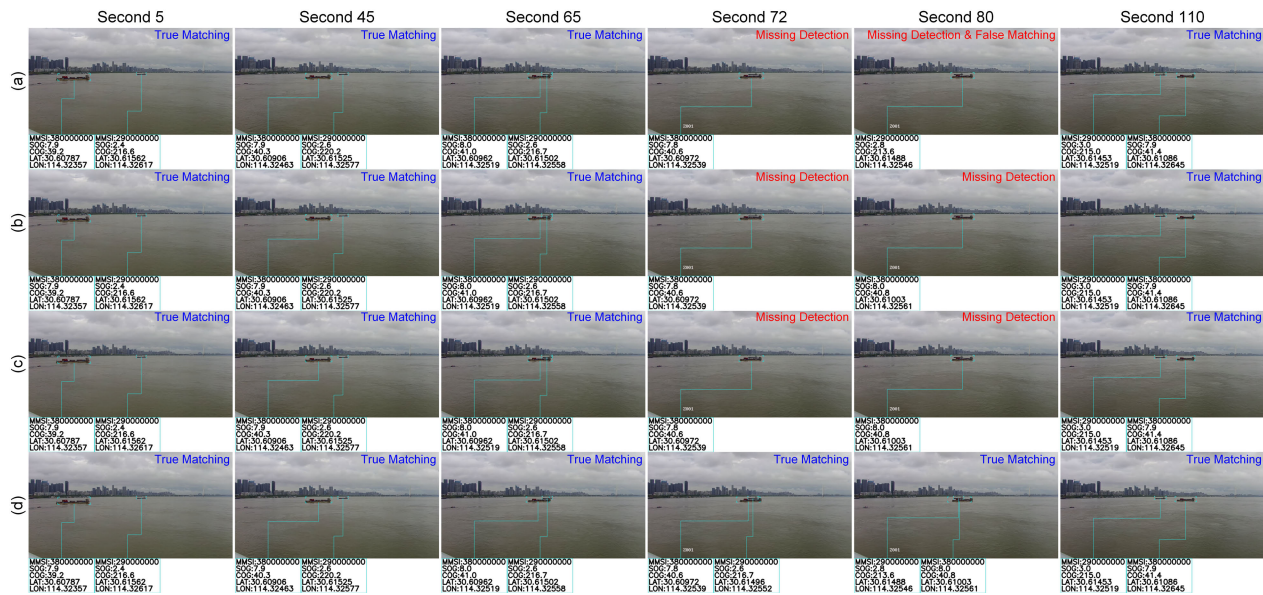


Fig. 5. Visual comparisons of fusion results on the dataset captured by the bridge region camera from Table II. From top to bottom: visual fusion results generated by (a) MSDF [9], (b) MMDST [8], (c) DeepSORVF without the anti-occlusion strategy, and (d) DeepSORVF, respectively.

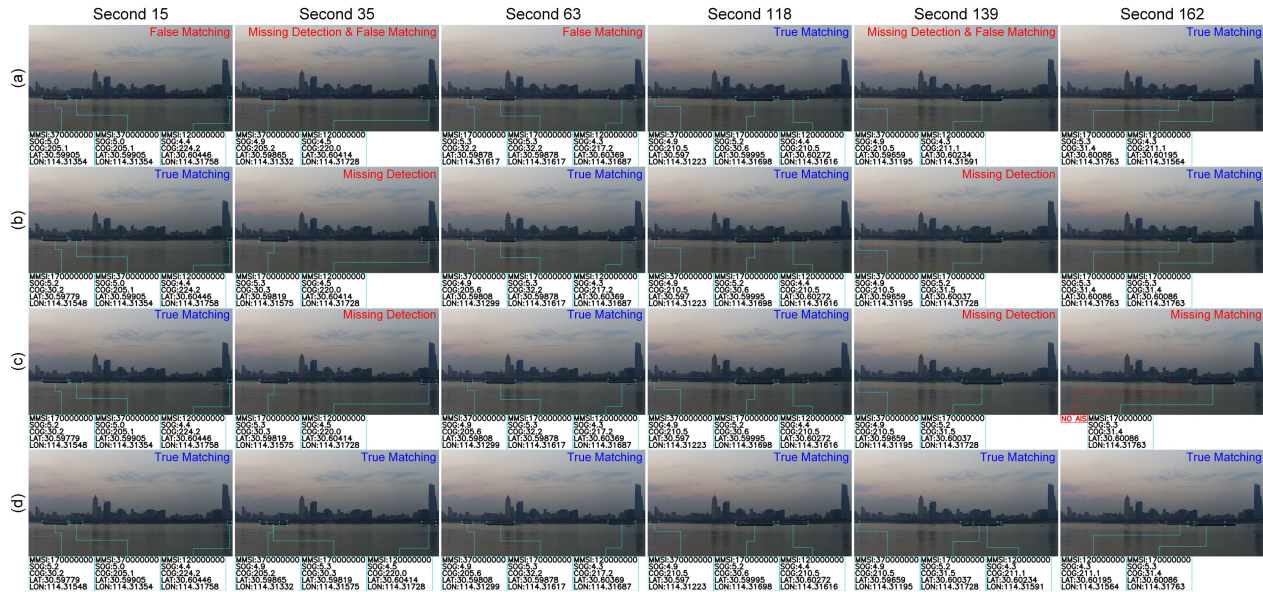


Fig. 6. Visual comparisons of fusion results on the dataset captured by the riverside camera from Table II. From top to bottom: visual fusion results generated by (a) MSDF [9], (b) MMDST [8], (c) DeepSORVF without the anti-occlusion strategy, and (d) DeepSORVF, respectively.

DeepSORVF (w/o) are unable to match the vessel identification information since the detector fails to identify the partially occluded target. For the data collected by the riverside, the visual vessels are often more severely occluded, resulting in the complete disappearance of target features. By analyzing Fig. 6, the MSDF, MMDST, and DeepSORVF (w/o) will produce more missing detection and false matching. It is worth mentioning that the vessel occlusion will also affect the trajectory feature extraction and cause the feature matching failure. In contrast, the proposed DeepSORVF with the anti-occlusion strategy has a more stable data fusion effect and is suitable for a variety of scenarios.

3) *Fusion Results on FVessel Dataset*: Our DeepSORVF is also used to process more data in the FVessel dataset and calculate the MOFA, IDP, IDR, and  $IDF_1$ . Table IV and Fig. 7 display the metric calculation results and the visualized fusion

results, respectively. It can be found that the proposed method has stable fusion performance. The fusion accuracy (MOFA) of our method is between 73.19% and 98.68%, and the average is 91.13%. In the evaluation of the other three metrics, our method also has a good performance. Through the comparison in Fig. 7, the results generated by our DeepSORVF are accurate and stable. The superiority of the proposed method benefits from the accurate prediction of the vessel bounding box by the anti-occlusion tracking method under the occlusion condition and the accurate matching based on the trajectory series.

### C. Influence of Data Fusion on Vessel Detection and Tracking

In our proposed method, the result of trajectory matching is fed as prior knowledge to the vessel detection and tracking



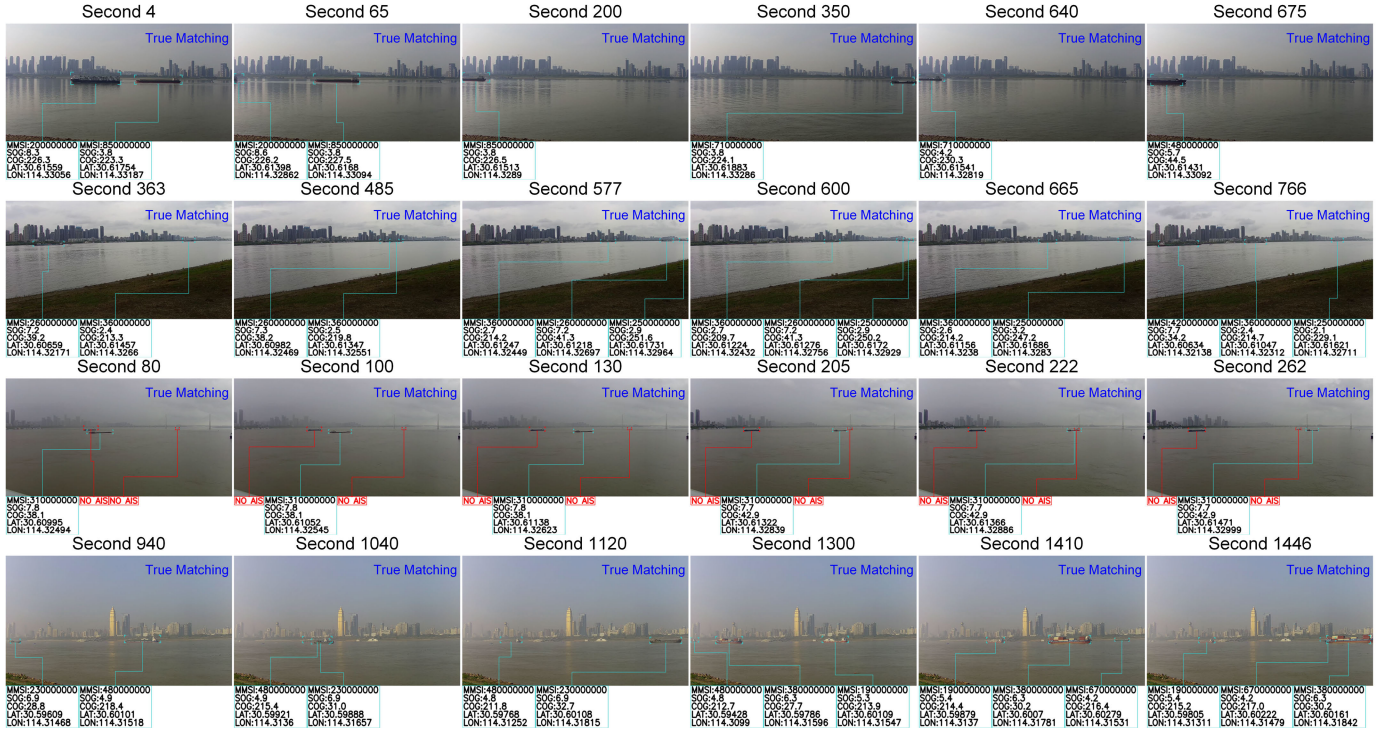


Fig. 7. Visual fusion results of our DeepSORVF on the FVessel dataset from Table I.

TABLE IV  
MOFA, IDP, IDR, AND IDF<sub>1</sub> RESULTS OF DATA FUSION FOR THE FVESSEL FROM TABLE I. (UNIT: %)

Video	MOFA ↑	IDP ↑	IDR ↑	IDF <sub>1</sub> ↑
video-01	79.94	89.35	90.76	90.05
video-02	73.19	83.27	91.60	87.23
video-03	96.45	99.23	97.20	98.20
video-04	98.08	99.45	98.63	99.03
video-05	89.19	93.46	95.91	94.67
video-06	91.17	96.04	95.08	95.56
video-07	96.81	99.59	97.21	98.39
video-08	82.28	99.64	82.58	90.31
video-09	98.45	100.00	98.45	99.22
video-10	88.74	90.42	99.26	94.63
video-11	97.66	99.29	98.36	98.83
video-12	95.45	99.06	96.36	97.69
video-13	84.82	94.82	89.72	92.20
video-14	93.10	97.82	95.22	96.50
video-15	95.88	97.19	98.74	97.96
video-16	98.68	100.00	98.68	99.33
video-17	90.02	93.80	96.39	95.08
video-18	74.49	83.57	92.72	87.91
video-19	96.62	98.31	98.31	98.31
video-20	96.74	98.66	98.07	98.36
video-21	76.43	87.03	89.82	88.40
video-22	96.82	99.35	97.45	98.39
video-23	94.71	98.91	95.77	97.31
video-24	94.70	98.34	96.33	97.32
video-25	91.49	97.66	93.73	95.66
video-26	97.44	99.11	98.32	98.72
Average	91.13	95.90	95.41	95.59

tasks at the next moment for promoting the video-based vessel trajectory extraction. To verify that our proposed data fusion method can improve vessel detection and tracking performance, we conduct massive experiments on ten clips. In particular, we select five different deep neural networks as detectors, i.e., Faster-RCNN [25], SSD [26], YOLOv4 [44], YOLOv5 [45], and YOLOX [30]. Each detector has two versions, i.e., “Detection” and “Detection + Data Fusion”. Furthermore, all methods only process data one per second.

1) *Evaluation Metric*: To evaluate the performance of vessel detection, we select the Precision and Recall as evaluation metrics. Let  $TP$ ,  $FP$ , and  $FN$  denote the number of the true positive, false positive, and false negative, the Precision and Recall can be given by

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (19)$$

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (20)$$

For vessel tracking, we tend to use MOTA as an evaluation metric, which can be defined as

$$\text{MOTA} = 1 - \frac{FP + FN + ID_s}{GT}, \quad (21)$$

where  $FP$ ,  $FN$ ,  $ID_s$ , and  $GT$  represent the numbers of the false positive, false negative,  $ID$  switch, and ground truth, respectively. Furthermore, we also adopt the IDP, IDR, and IDF<sub>1</sub> metrics. Theoretically, better detection results have higher Precision and Recall, and better tracking results have higher MOTA, IDP, IDR, and IDF<sub>1</sub>.

2) *Vessel Detection and Tracking on Ten Clips*: Table V compares the detection Precision and Recall of various detectors on ten clips. Due to the mutual occlusion between the targets, some vessel characteristics are easily hidden by another vessel. Therefore, detectors often suffer from missing detection, resulting in higher  $FN$  and poorer Recall. In most cases, detectors are prone to produce false detection boxes in vessel encounter regions due to the overlapping of multiple vessel features. These false detection boxes will produce higher  $FP$  and poorer Precision. In contrast, the proposed anti-occlusion method based on data fusion results can improve the performance of various detectors.



TABLE V  
PRECISION AND RECALL RESULTS OF VESSEL DETECTION  
FOR THE TEN CLIPS FROM TABLE II. (UNIT: %)

Detector	Data Fusion	Precision $\uparrow$	Recall $\uparrow$
Faster-RCNN [25]	✗	89.04	90.74
	✓	<b>90.03</b>	<b>96.48</b>
SSD [26]	✗	94.98	88.14
	✓	<b>95.02</b>	<b>95.62</b>
YOLOv4 [44]	✗	96.43	76.97
	✓	<b>96.74</b>	<b>84.85</b>
YOLOv5 [45]	✗	98.78	83.70
	✓	<b>99.47</b>	<b>91.29</b>
YOLOX [30]	✗	98.91	94.13
	✓	<b>99.20</b>	<b>99.40</b>

TABLE VI  
MOTA, IDP, IDR, AND IDF<sub>1</sub> RESULTS OF VESSEL TRACKING  
FOR THE TEN CLIPS FROM TABLE II. (UNIT: %)

Detector	Data Fusion	MOTA $\uparrow$	IDP $\uparrow$	IDR $\uparrow$	IDF <sub>1</sub> $\uparrow$
Faster-RCNN [25]	✗	75.19	68.03	69.00	68.43
	✓	<b>84.43</b>	<b>81.81</b>	<b>87.95</b>	<b>84.58</b>
SSD [26]	✗	81.36	84.49	78.77	81.14
	✓	<b>89.69</b>	<b>90.03</b>	<b>91.03</b>	<b>90.13</b>
YOLOv4 [44]	✗	70.92	71.61	56.66	63.08
	✓	<b>80.50</b>	<b>80.44</b>	<b>70.18</b>	<b>74.77</b>
YOLOv5 [45]	✗	81.72	88.68	75.08	81.12
	✓	<b>90.40</b>	<b>94.48</b>	<b>86.50</b>	<b>90.05</b>
YOLOX [30]	✗	92.61	89.80	85.36	87.50
	✓	<b>98.61</b>	<b>97.72</b>	<b>97.92</b>	<b>97.82</b>

To compare the tracking performance, Table VI further shows the MOTA, IDP, IDR, and IDF<sub>1</sub> results of various detectors on ten clips. In contrast, the proposed data fusion method can significantly improve the tracking performance of all five detectors and reduce the number of missing and false detection. The performance improvement benefits from the proposed anti-occlusion method based on data fusion results. The proposed method can achieve more stable vessel tracking during the occlusion.

#### D. Running Time Analysis

The time complexity of the proposed method is a critical metric which directly determines whether it can be used in actual engineering. In this work, we only process the data once per second to ensure practicability. Therefore, we are unable to use the frame per second (FPS) as an evaluation metric. Meanwhile, since the proposed method considers trajectory features, the time complexity is also related to the number and length of AIS- and video-based vessel trajectories. Consequently, it is also inaccurate to calculate the running time of a single image. Finally, we compute the processing time of one-second data for ten clips in Table II. The processing time of our method for each clip is shown in Fig. 8 and Table VII. It can be seen that our DeepSORVF has low time complexity and high practicability. It can process one second of data in 0.175-0.500 seconds and 0.2562 seconds on average.

#### E. Discussion

Although our proposed method adopts the prior knowledge-driven anti-occlusion tracking method and trajectory matching

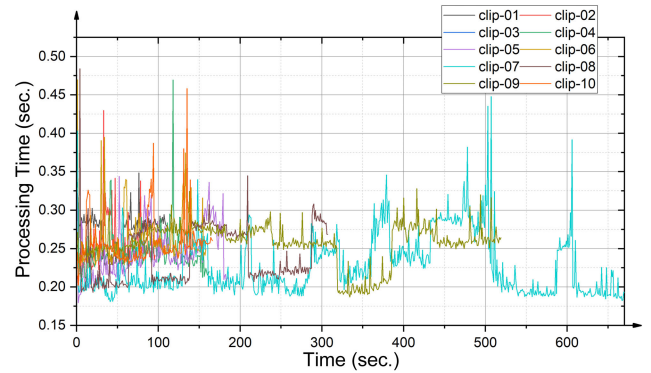


Fig. 8. Processing time of one-second data on the ten clips from Table II.

TABLE VII  
PROCESSING TIME OF ONE-SECOND DATA (MEAN  $\pm$  STD) ON  
THE TEN CLIPS FROM TABLE II (UNIT: Sec.)

Video	Video Length	Processing Time
clip-01	01m51s	0.2721 $\pm$ 0.0230
clip-02	01m36s	0.2526 $\pm$ 0.0323
clip-03	03m42s	0.2998 $\pm$ 0.0413
clip-04	02m42s	0.2462 $\pm$ 0.0289
clip-05	03m05s	0.2444 $\pm$ 0.0321
clip-06	02m38s	0.2642 $\pm$ 0.0344
clip-07	11m10s	0.2279 $\pm$ 0.0391
clip-08	05m07s	0.2404 $\pm$ 0.0364
clip-09	08m39s	0.2569 $\pm$ 0.0256
clip-10	02m46s	0.2579 $\pm$ 0.0304
Average	–	0.2562 $\pm$ 0.0323

TABLE VIII  
MOFA (%) AND MOFP RESULTS OF DATA FUSION FOR THE CLIP-02  
AND CLIP-03 FROM TABLE II. DEEPSORVF (W/O) REPRESENTS OUR  
DEEPSORVF WITHOUT THE ANTI-OCCLUSION STRATEGY

Video	Methods	MOFA $\uparrow$	MOFP $\downarrow$
clip-02	DeepSORVF (w/o)	87.77	<b>0.110</b>
	DeepSORVF	<b>99.47</b>	0.140
clip-03	DeepSORVF (w/o)	90.53	<b>0.241</b>
	DeepSORVF	<b>94.67</b>	0.252
Average	DeepSORVF (w/o)	88.92	<b>0.175</b>
	DeepSORVF	<b>97.02</b>	0.196

method to effectively improve the accuracy of data fusion, our method still has two key limitations that require further investigation and improvement in future research.

1) *Precision of Predicted Boxes Under the Occlusion:* We use the multiple object fusion accuracy (MOFA) and multiple object fusion precision (MOFP) as evaluation metrics, where the MOFP is a variant of the multiple object tracking precision (MOTP) [42] in the data fusion. The MOFP can be given by

$$\text{MOFP} = \frac{\sum_t D_{\text{mmsi}}^{t,i}}{\sum_t N_{\text{mmsi}}^t}, \quad (22)$$

where  $D_{\text{mmsi}}^{t,i}$  denotes the distance of the  $i$ -th MMSI matching pair in the  $t$ -th second,  $N_{\text{mmsi}}^t$  is the number of matches in the  $t$ -th second. Theoretically, a better fusion effect has higher MOFA and lower MOFP.

Using clip-02 and clip-03 as examples, we compute their MOFA and MOFP. As shown in Table VIII, it can be found that the proposed anti-occlusion tracking method can significantly improve the accuracy of data fusion by comparing the MOFA. However, the DeepSORVF is slightly inferior

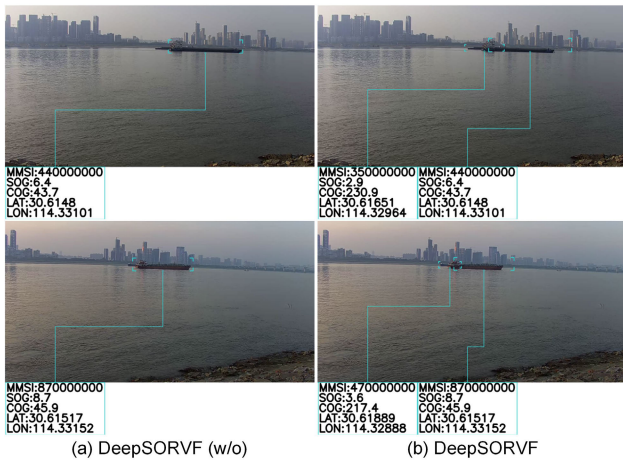


Fig. 9. Visual comparisons of fusion results on the dataset from Table II. DeepSORVF (w/o) represents our DeepSORVF without the anti-occlusion strategy.

to the DeepSORVF (w/o) without the anti-occlusion strategy in the bounding box localization precision evaluated by the MOFP. The more intuitive comparisons before and after the use of the anti-occlusion strategy are illustrated in Fig. 9. Our DeepSORVF with the anti-occlusion strategy can predict the vessel position and accurately match the occluded vessel information. However, the predicted bounding boxes still have some degree of bias in complex occlusion conditions. This deviation is mostly attributable to the inaccurate estimation of AIS and visual motion characteristics. When a vessel travels away from the camera, for instance, its visual movement speed generally slows, and the object gets smaller. To further improve the vessel anti-occlusion performance, our future work will take into account the changing features of the moving vessels in the visual data.

2) *Usage in Congested Waterway*: It is challenging to practically apply our method under highly congested waterway conditions. Firstly, the large number of vessels and complex trajectories result in decreased accuracy of anti-occlusion tracking and data matching results. Secondly, highly congested waterways mean a large number of vessels need to be monitored and tracked, increasing the amount of data and computing resources required by the system. Therefore, our future research will focus on the design of high-accuracy and low-computation methods to overcome these limitations and ensure accurate vessel detection and tracking, as well as high reliability of data fusion.

## V. CONCLUSION

In this paper, we proposed a deep learning-based simple online and real-time vessel data fusion method (named DeepSORVF). The DeepSORVF could pair the vessel features of AIS with visual targets. Due to the fact that reciprocal occlusion between vessel targets may readily interfere with video-based trajectory extraction, we suggested a prior knowledge-driven anti-occlusion tracking method. Meanwhile, a novel asynchronous trajectory matching method was designed for robust data fusion. Comprehensive experiments on vessel detection, vessel tracking, data fusion, and running

time analysis have demonstrated the superior performance of our DeepSORVF on the newly-developed FVessel dataset.

## ACKNOWLEDGMENT

The authors deeply thank Jianlong Su from the School of Computer and Artificial Intelligence, Wuhan University of Technology, who kindly performs the data acquisition and processing tasks.

## REFERENCES

- [1] D. D. Bloisi, F. Previtali, A. Pennisi, D. Nardi, and M. Fiorini, "Enhancing automatic maritime surveillance systems with visual information," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 824–833, Apr. 2017.
- [2] X. Zhang, X. Fu, Z. Xiao, H. Xu, and Z. Qin, "Vessel trajectory prediction in maritime transportation: Current approaches and beyond," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 19980–19998, Nov. 2022.
- [3] D. Y. Kim, B. Ristic, R. Guan, and L. Rosenberg, "A Bernoulli track-before-detect filter for interacting targets in maritime radar," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 3, pp. 1981–1991, Jun. 2021.
- [4] J. Feng, B. Li, L. Tian, and C. Dong, "Rapid ship detection method on movable platform based on discriminative multi-size gradient features and multi-branch support vector machine," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 1357–1367, Feb. 2022.
- [5] J. Chen, Q. Hu, R. Zhao, P. Guojun, and C. Yang, "Tracking a vessel by combining video and AIS reports," in *Proc. 2nd Int. Conf. Future Gener. Commun. Netw.*, Dec. 2008, pp. 374–378.
- [6] W. Man and L. Zhiyong, "The information fusion based on AIS and video data," in *Proc. 5th Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT)*, Dec. 2016, pp. 336–339.
- [7] Y. Lu et al., "Fusion of camera-based vessel detection and AIS for maritime surveillance," in *Proc. 26th Int. Conf. Autom. Comput. (ICAC)*, Sep. 2021, pp. 1–6.
- [8] Z. Huang, Q. Hu, Q. Mei, C. Yang, and Z. Wu, "Identity recognition on waterways: A novel ship information tracking method based on multimodal data," *J. Navigat.*, vol. 74, no. 6, pp. 1336–1352, Nov. 2021.
- [9] R. W. Liu et al., "Intelligent edge-enabled efficient multi-source data fusion for autonomous surface vehicles in maritime Internet of Things," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 3, pp. 1574–1587, Sep. 2022.
- [10] J. Qu, Y. Guo, Y. Lu, F. Zhu, Y. Huan, and R. W. Liu, "Intelligent maritime surveillance framework driven by fusion of camera-based vessel detection and AIS data," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 2280–2285.
- [11] J. Qu, R. W. Liu, Y. Guo, Y. Lu, J. Su, and P. Li, "Improving maritime traffic surveillance in inland waterways using the robust fusion of AIS and visual data," *Ocean Eng.*, vol. 275, May 2023, Art. no. 114198.
- [12] K. El-Darymli, P. McGuire, D. Power, and C. Moloney, "Target detection in synthetic aperture radar imagery: A state-of-the-art survey," *J. Appl. Remote Sens.*, vol. 7, no. 1, Mar. 2013, Art. no. 071598.
- [13] W. Zhou and P. Chen, "A deep attention mechanism method for maritime salient ship detection in complex sea background," *Optoelectronics Lett.*, vol. 17, no. 7, pp. 438–443, Jul. 2021.
- [14] R. W. Liu, W. Yuan, X. Chen, and Y. Lu, "An enhanced CNN-enabled learning method for promoting ship detection in maritime surveillance system," *Ocean Eng.*, vol. 235, Sep. 2021, Art. no. 109435.
- [15] X. Chen, X. Xu, Y. Yang, Y. Huang, J. Chen, and Y. Yan, "Visual ship tracking via a hybrid kernelized correlation filter and anomaly cleansing framework," *Appl. Ocean Res.*, vol. 106, Jan. 2021, Art. no. 102455.
- [16] D.-W. Gao, Y.-S. Zhu, J.-F. Zhang, Y.-K. He, K. Yan, and B.-R. Yan, "A novel MP-LSTM method for ship trajectory prediction based on AIS data," *Ocean Eng.*, vol. 228, May 2021, Art. no. 108956.
- [17] Z. Xiao, X. Fu, L. Zhang, and R. S. M. Goh, "Traffic pattern mining and forecasting technologies in maritime traffic service networks: A comprehensive survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1796–1825, May 2020.
- [18] D. K. Prasad, C. K. Prasad, D. Rajan, L. Rachmawati, E. Rajabally, and C. Quek, "Object detection in a maritime environment: Performance evaluation of background subtraction methods," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1787–1802, May 2019.

- [19] W.-C. Hu, C.-Y. Yang, and D.-Y. Huang, "Robust real-time ship detection and tracking for visual surveillance of cage aquaculture," *J. Vis. Commun. Image Represent.*, vol. 22, no. 6, pp. 543–556, Aug. 2011.
- [20] D. D. Bloisi, A. Pennisi, and L. Iocchi, "Background modeling in the maritime domain," *Mach. Vis. Appl.*, vol. 25, no. 5, pp. 1257–1269, Jul. 2014.
- [21] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3446–3456, Sep. 2010.
- [22] Y. Zhang, Q.-Z. Li, and F.-N. Zang, "Ship detection for visual maritime surveillance from non-stationary platforms," *Ocean Eng.*, vol. 141, pp. 53–63, Sep. 2017.
- [23] X. Chen, S. Wang, C. Shi, H. Wu, J. Zhao, and J. Fu, "Robust ship tracking via multi-view learning and sparse representation," *J. Navigat.*, vol. 72, no. 1, pp. 176–192, Jan. 2019.
- [24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.
- [25] 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, Jun. 2017.
- [26] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. ECCV*, 2016, pp. 21–37.
- [27] Z. Li and F. Zhou, "FSSD: Feature fusion single shot multibox detector," 2017, *arXiv:1712.00960*.
- [28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
- [29] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
- [30] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
- [31] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [32] Z. Shao, L. Wang, Z. Wang, W. Du, and W. Wu, "Saliency-aware convolution neural network for ship detection in surveillance video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 781–794, Mar. 2020.
- [33] Z. Chen, D. Chen, Y. Zhang, X. Cheng, M. Zhang, and C. Wu, "Deep learning for autonomous ship-oriented small ship detection," *Saf. Sci.*, vol. 130, Oct. 2020, Art. no. 104812.
- [34] Y. Xi, Y. Wang, N. Wang, and X. Gao, "An enhanced SiamMask network for coastal ship tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5612011.
- [35] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.
- [36] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 666–673.
- [37] B. Bovcon, J. Muhovic, D. Vranac, D. Mozetič, J. Perš, and M. Kristan, "MODS—A USV-oriented object detection and obstacle segmentation benchmark," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13403–13418, Aug. 2022.
- [38] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.
- [39] M. Müller, "Dynamic time warping," *Inform. Retr. Music Motion*, pp. 69–84, Jan. 2007.
- [40] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intell. Data Anal.*, vol. 11, no. 5, pp. 561–580, Oct. 2007.
- [41] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.
- [42] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, Nov. 2008.
- [43] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV*, 2016, pp. 17–35.
- [44] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
- [45] G. Jocher et al., "Ultralytics/YOLOv5: V5.0—YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations," Apr. 2021.



**Yu Guo** (Member, IEEE) received the B.E. degree in naval architecture and marine engineering from the School of Transportation, Wuhan University of Technology, Wuhan, China, in 2021. He is currently pursuing the Ph.D. degree in traffic information engineering and control with the School of Navigation, Wuhan University of Technology. His research interests include computer vision, machine learning, and intelligent navigation systems. He has received the Best Paper Award from the AAAI-22 Workshop: AI for Transportation.



**Ryan Wen Liu** (Member, IEEE) received the B.Sc. degree (Hons.) in information and computing science from the Department of Mathematics, Wuhan University of Technology, Wuhan, China, in 2009, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2015. He was a Visiting Scholar with the Agency for Science, Technology and Research, Singapore. He is currently a Professor with the School of Navigation, Wuhan University of Technology. His research interests include computer vision, data mining, and intelligent transportation.



**Jingxiang Qu** received the B.E. degree from the School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang, China, in 2021. He is currently pursuing the M.S. degree in traffic information engineering and control with the School of Navigation, Wuhan University of Technology, Wuhan, China. His research interests include intelligent transportation systems, machine learning, and computer vision.



**Yuxu Lu** received the B.E. degree in navigation technology from the School of Navigation, Wuhan University of Technology, Wuhan, China, in 2020, where he is currently pursuing the M.S. degree in navigation and information engineering. His research interests include computer vision, machine learning, and intelligent navigation systems. He has received the Excellent Oral Presentation from the 11th International Conference on Machine Learning and Computing (ICMLC 2019).



**Fenghua Zhu** (Senior Member, IEEE) received the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2008. He is currently an Associate Professor with the State Key Laboratory of Multimodal Artificial Intelligence Systems, China. His research interests include artificial transportation systems and parallel transportation management systems.



**Yisheng Lv** (Senior Member, IEEE) received the B.E. and M.E. degrees from the Harbin Institute of Technology, Harbin, China, in 2005 and 2007, respectively, and the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 2010. He is currently a Professor with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. His research interests include AI for ITS, intelligent vehicles, and parallel traffic management and control systems. He is an Associate Editor of the