Personnel Trajectory Extraction From Port-Like Videos Under Varied Rainy Interferences

Xinqiang Chen, *Member, IEEE*, Chenxin Wei, Yang Yang, Lijuan Luo, Salvatore Antonio Biancardo, and Xiaojun Mei, *Member, IEEE*

***Abstract*— Large-scale deployed cameras in the automated container terminal (ACT) area helps on-site staff better identify unexpected yet emergency events by monitoring port personnel trajectories. Rainy weather is *a* common yet typical problem which may significantly deteriorate trajectory extraction performance. To tackle the problem, the study proposes an ensemble framework to extract personnel trajectory from port-like surveillance videos under varied rainy weather scenarios. Firstly, the proposed framework learns fine-grained personnel features with the help of the object query and transformer encoder-decoder module from the input port-like image sequences, and thus obtains port personnel locations from the input low-visibility images. Secondly, the personnel positions are further associated in a frame-by-frame manner with the help of neighboring kinematic movement information and feature information. Finally, a memory mechanism is introduced in the proposed framework to suppress personnel trajectory discontinuity outlier. In that manner, we can obtain accurate yet consistent personnel trajectories, and each person is assigned with a unique ID. We verified the proposed model performance on three port-like rainy videos involving with interferences of rain, rain streak and fog. Experimental results show that the proposed port personnel trajectory extraction framework can obtain satisfied performance considering that the average multi-target accuracy (MOTA), the average value of judging the same target (IDF_1), average recall rate (IDR) and average precision (IDP) were larger than 92%.**

***Index Terms*— Personnel trajectory extraction, rainy interference, ensemble transformer framework, memory mechanism, automated container terminal.**

I. Introduction

THE automated container terminal (ACT) has undergone rapid development along with evolution of artificial intelligence techniques [1], [2]. It is crucial to quickly and accurately identify personnel locations from port videos for the purpose of early-warning emergency or abnormal events. It is noted that rainy weather is common in real-world port activities, which may impose negative impact on the trajectory exploitation related tasks [3]. Port personnel trajectory extraction performance under varied rainy weather may fail to obtain satisfactory results due to potential personnel visual feature deterioration caused by rain (e.g., raindrops, rain streaks) [4], [5]. Port videos captured by a camera attached with rain drops were similar to those taken by a fisheye camera [6]. Raindrops play the role of blurry mask, which obscure the port image sequence quality. It is also noted that rain and fog in the rainy weather condition can significantly degrade image quality due to that rain and fog-related pixels may contaminate port person features [7], [8], [9].

Many focuses have been paid to explore kinematic spatial-temporal data (trajectory, speed, etc.) from varied surveillance videos captured under good weather conditions. Chen et al., extracted vehicle speed data from drone videos with an ensemble detection framework [10]. Sousa et al., conducted a comprehensive review of state-of-art methods about vehicular trajectory extraction and representation [11]. It is found that object trajectories can be successfully extracted from port video streams while weather conditions are good [12], [13]. Object detection with segmentation models is also conducted to detect unknown targets in images. For instance, K. Sirohi et al., propose a novel top-down evidential panoptic segmentation network to identify objects via the help of panoptic fusion module [14]. Li et al., implemented unfamiliar object detection task by integrating region proposal network and the radial basis function network [15].

Previous studies demonstrate that imaging spatial-temporal data can be obtained against adverse weather interference via support of multi-sensor data fusion mechanism [16], [17]. Bai et al., obtained high-fidelity target motion information by fusing radar and camera data using Gaussian mixture model [18]. Liu et al., introduced a two-layer feature model to exploit discriminant features from thermal infrared images [19]. The main weakness is that multi-sensor data fusion-based models may need to implement coordinate registration to map reference coordinates for different data sources into same coordinate system [20]. Rain, magnetic interference from yard crane and corrosive environmental condition may reduce the above-mentioned model performance in the port personal trajectory extraction task [21], [22].

To address the issue, we propose a novel end-to-end deep learning framework for extracting port personnel trajectories.

Manuscript received 3 January 2023; revised 17 June 2023 and 5 November 2023; accepted 16 December 2023. Date of publication 13 February 2024; date of current version 2 July 2024. This work was supported by the National Natural Science Foundation of China under Grant 52331012, Grant 52102397, Grant 52071200, Grant 72101157, Grant 71942003, and Grant 52201401. The Associate Editor for this article was Z. Duric. *(Corresponding author: Lijuan Luo.)*

Xinqiang Chen and Chenxin Wei are with the Institute of Logistics Science and Engineering, Shanghai Maritime University, Shanghai 201306, China (e-mail: xqchen@shmtu.edu.cn; 202130510052@stu.shmtu.edu.cn).

Yang Yang is with the School of Transportation Science and Engineering, Beihang University, Beijing 100083, China (e-mail: yangphd@buaa.edu.cn).

Lijuan Luo is with the Key Laboratory of Brain-Machine Intelligence for Information Behavior (Ministry of Education and Shanghai), School of Business and Management, Shanghai International Studies University, Shanghai 201620, China (e-mail: luolijuan@shisu.edu.cn).

Salvatore Antonio Biancardo is with the School of Polytechnic and Basic Sciences, University of Naples Federico II, 80138 Naples, Italy (e-mail: Salvatoreantonio.biancardo@unina.it).

Xiaojun Mei is with the Merchant Marine College, Shanghai Maritime University, Shanghai 201306, China (e-mail: xjmei@shmtu.edu.cn).

Digital Object Identifier 10.1109/TITS.2023.3346473

This framework consists of personnel detection from port videos and personnel ID association. The primary academic contributions of the study can be summarized as follows:

- we propose a novel ensemble framework to extract port personnel trajectories under rainy weather interference. We formulate the port personnel detection problem into an ensemble prediction task, which mitigates the disadvantages of non-maximum suppression, anchor link misconnection, etc.
- we introduce a memory module into the straightforward online tracking model with a deep association metric module to suppress rain and occlusion imaging interferences in port videos.
- we verified that the proposed framework achieved satisfactory results in port-like videos under varied rainy interferences. (i.e., raindrop, rain streak, hybrid interference of rain and fog).

II. Related Works

Trajectory extraction has become a hot topic in the transportation community along with the artificial intelligence technique development. A bunch of studies have been conducted to exploit trajectories from varied data sources. In this section, we focus on reviewing the state-of-the-art methods for trajectory extraction with deep learning and multi-sensor approaches.

A. Trajectory Extraction From Video Data

Multi-object tracking algorithms have been proposed for good weather conditions [11], [12]. It is noted that previous trajectory extraction related studies are conducted under good weather conditions [7]. Some scholars try to extract trajectory from low visibility videos with image restoration and enhancement. M. Hassaballah et al., utilized a visibility enhancement scheme to preprocess video images, and obtained vehicle trajectory by proposing a multi-scale deep convolution tracking method [23]. Quan et al., proposed a complementary cascaded network framework to remove rain interference in video data for autonomous driving scenarios [24]. Wu et al., proposed a class encoder framework based on an adaptive mixup operation and a dynamic feature enhancement module to achieve video image defogging [21]. Similar studies can also be found in [25].

B. Trajectory Extraction From Thermal Images

Thermal imager plays an important role in various monitoring scenarios due to its insensitivity to visibility. The thermal camera can be deployed in low visibility environments such as nighttime and foggy conditions [26]. KRIŠTO et al., utilized visual technology and the difference in thermal image features to implement object detection on thermal imaging data using you only look once (YOLOv3) model [27]. M. P. Muresan et al., developed a Siamese network to implement real-time pedestrian detection and tracking from thermal images with the help of original edge-based descriptor and data association method [28]. Yuan et al., proposed an efficient thermal infrared target tracking method by utilizing a spatial-temporal memory network model and an alignment matching module to model and spatially correct information in the infrared target tracking scenario [29].

C. Trajectory Extraction From Multiple Data Sources

Multi-sensor data fusion scheme has shown its superiority in tackling the challenges of clutter interference, object occlusion, and limited sensor deployment in complex scenes. B. Iepure et al., proposed a novel object tracking method with data collected from thermal sensors, optical sensors and millimeter-wave sensors [30]. Ouyang et al., proposed a novel SacadeFork model to accurately detect vehicle and pedestrian under different scenarios by fusing image and LiDAR point cloud data [31]. Some attentions are also paid to exploit trajectory data from high-resolution satellite and radar images [32].

In sum, the above-mentioned models can extract trajectory in complex environments, which may fail to obtain accurate personnel trajectories from port-like videos due to the following reasons: (1) the video-based trajectory extraction models may require image restoration procedure. It is found that the step is computer hardware demanding, which can be hardly deployed in the real-world port personal trajectory extraction task; (2) thermal image quality is easily affected by the port environment, while the image contrast may be quite low (i.e., object resolution in the images may be too low to be identified); (3) it was difficult to deploy multiple yet varied sensors in port area due to port operation safety and sensor heath status maintenance, etc. We aim to extract high-fidelity port personnel from videos with hybrid rainy interferences without image restoration procedure. The proposed model accurately identifies port personnel from rainy-polluted images through a multi-head attention mechanism, and accurately realizes trajectory extraction.

III. Proposed Solution

A. Framework Overview

The proposed port personnel trajectory extraction framework mainly consists of personnel detection and imaging position data association (see figure 1). Firstly, we employ the convolution neural network (CNN) network to extract object features from the collected port videos. We also introduce a multi-head attention mechanism for context interaction with the support of local and global feature. Besides, the object query random variable is developed to introduce small biases into the object query learning iteration procedure. The proposed object query random variable integrates object feature and position into a model to enhance port personnel extraction accuracy under adverse weather interference. Secondly, we propose an improved DeepSort algorithm with memory module to store (and recover) port personnel trajectory data. The memory mechanism helps the proposed framework repair trajectory data loss (and ID switch) like outliers by storing historical port personnel position ID and feature information.

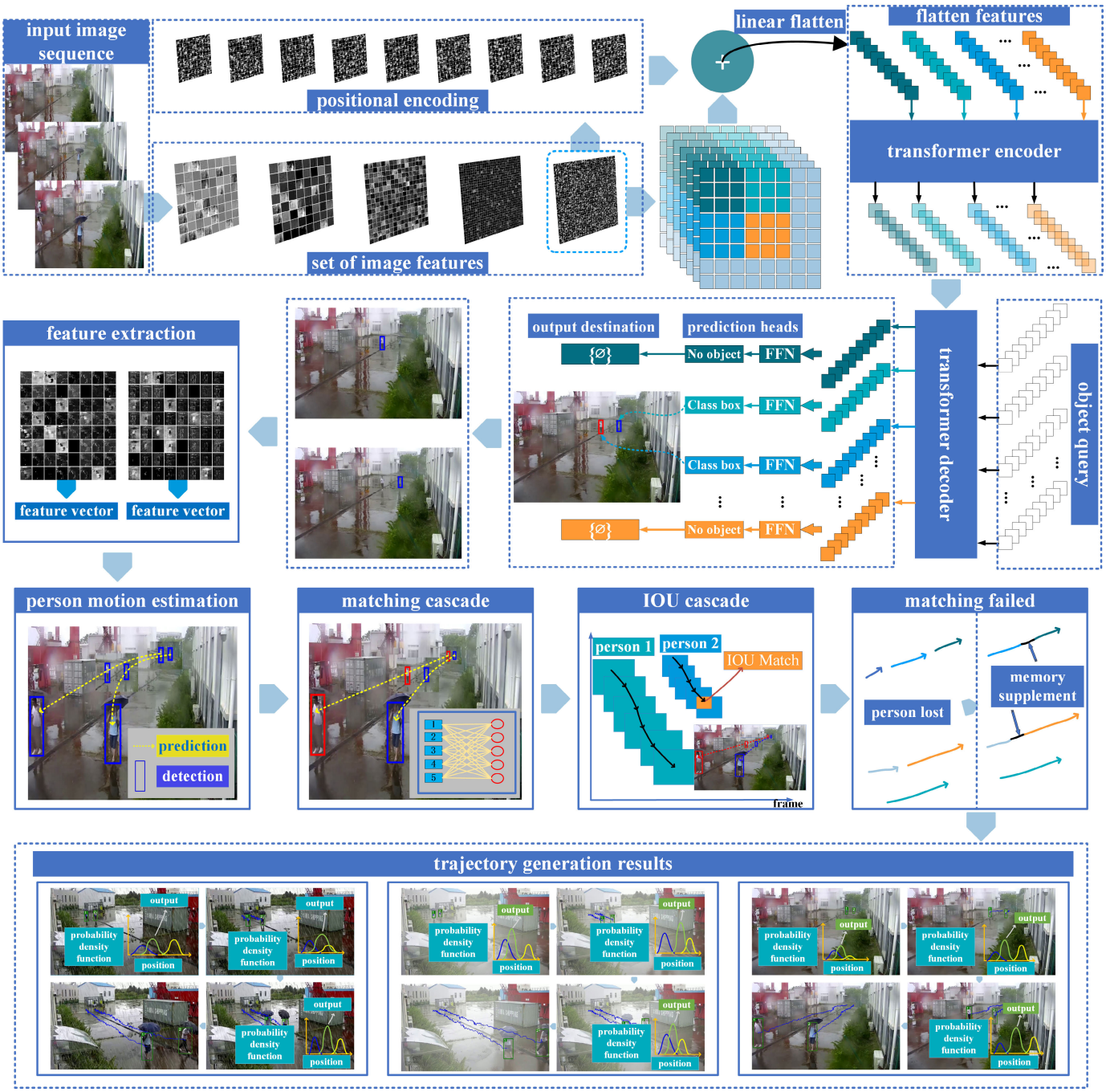


Fig. 1. Schematic overview for port personnel trajectory extraction under low-visibility port environments.

B. Port Personnel Detection With Transformer Detector

Port surveillance videos collected under adverse weather condition and complicated port area (i.e., varied imaging interference) challenges personnel detection model performance. We extract environmental features from port image sequences with the help of traditional CNN neural network, and thus obtain a new yet distinct feature map. Furthermore, we propose an end-to-end low-visibility personnel detector to obtain personnel ID from port surveillance videos. The main advantage of the proposed port personnel detector is that the model formulates the detection task into an ensemble prediction problem.

Stage 1: Personnel detection with transformer encoder

The feature map $\mathcal{L}_f$ obtained by the CNN is compressed into a new feature map $\mathcal{L}_0$ from dimension C to dimension d with a 1×1 convolution kernel. The feature map $\mathcal{L}_0$ is flattened to obtain a $d\times Row \times Col$ feature map considering that the transformer module input is a sequence vector. The feature map is further unfold into a one-dimensional vector (with size $Row\times Col$) to efficiently predict bounding boxes (Bbox). The feature vector and encoded spatial position are concatenated as the input to the transformer encoder to mitigate the weakness of sequence order insensitivity. We calculate the fixed position encoding of the two measurements with Eq. (1) and (2) to

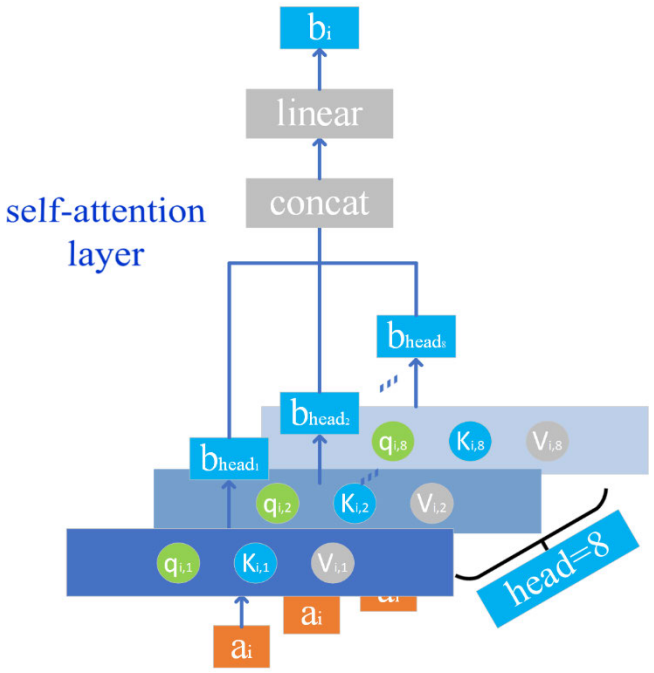


Fig. 2. Port personnel detection oriented multi-head attention structure via encoder and decoder mechanism.

specify image information in both of X and Y axis. Note that the procedure is integrated into each attention layer.

$$Epos_{(Sep,2n)} = \sin\left(\frac{Sep}{10000^{2n/dim}}\right) \tag{1}$$

$$Epos_{(Sep,2n+1)} = \cos\left(\frac{Sep}{10000^{2n/dim}}\right) \tag{2}$$

where Sep represents image sequence position, and dim represents the vector dimension. The 2n and 2n+1 represent the even dimension and odd dimension in the dim, respectively. It can be seen from the nature of the triangle function that $Sep+k$ can be calculated in the Sep position at each position, and position encode in each dimension is unique from each other.

Each encoding layer in the transformer includes a multi-headed self-attention module and a feed-forward network (FFN). Note that multi-headed self-attention module in our study consists of 8 heads, which can efficiently focus on fine-grained and important features of port personnel from low-visibility images. The proposed multi-head attention structure is shown in figure 2. The multi-head attention module splits the input feature vector into 8 heads (i.e., 8 feature sub-vector). In that way, we employ self-attention mechanism on each head to holistically exploit features at different levels from input port image sequences. Then, the explored features from sub-vectors are aggregated into a global feature map.

The multi-head attention assigns attention coefficients to each head, and thus enables the selective weighting and integration of feature information from different heads. The operation can efficiently retrieve personal feature maps from various locations in the port video clips. The attention mechanism is formulated with Eq. (3), (4) and (5). In that manner, we can obtain holistic global feature representation with the feedforward network.

$$Att\,(m,e,r) = Softmax\left(\frac{m\cdot e^{T}}{\sqrt{n_k}}\right) r \tag{3}$$

$$hd_i = Att\left(m_i w_i^m, e_i w_i^e, r_i w_i^r\right) \tag{4}$$

$$Mhd\,(m,e,r) = Concat\,(hd_1, hd_2 \ldots\ldots hd_8)\, w^o \tag{5}$$

where $Att\,(m,e,r)$ represents the weighted feature vector, m represents the weight of the query vector matrix, e represents the weight of the key vector matrix, r represents the weight of the value vector matrix, $\sqrt{n_k}$ represents a scaling factor that keeps the gradient stable, n_k represents the dimension of the input. hd_i represents the i-th attention head, $Mhd\,(m,e,r)$ represents weighted integration. w_i^m, w_i^e, w_i^r, w^o are all weight matrices, T represents the transpose matrix.

Stage 2: Personnel detection of transformer decoder

Both of decoder and encoder architectures are similar to each other, while the encoder architecture consists of an additional multi-head attention module. The input of decoder multi-head attention model is global feature information of port-like regions and object query (the fine-grained features of port personnel) output by the previous layer encoder (or the previous layer decoder). Note that the module in the proposed framework can simultaneously decode N objects. Due to the position invariant constraint of the encoder, to ensure the correspondence of the features of port-like personnel, the same position encoding is added to each attention layer in the decoder. The object query in the decoder consists of N d-dimensional vectors, similar to the spatial position encoding in the encoder, which are learned position encodings. After being processed by the encoder, the positional embedding carries regional features of the locations of low-visibility personnel that each query focuses on. The object query, which carries information about the objects in the image, is initially unaware of the objects in the image before entering the decoder. Therefore, it is set as a random variable. During model training, these object queries can cover the entire image as evenly as possible, which allows for better perception of the global image features of port-like areas and the context inference of personnel features.

The object query interacts with the encoder at the corresponding position through self-attention and cross-attention learning. In this process, the self-attention query distinguishes between foreground (port-like personnel) and background targets (containers, AGVs, etc.) in each frame of the image. Meanwhile, cross-attention allows each query to extract edge feature information of low-visibility personnel from the image based on their respective areas of interest. Specifically, each query pays more attention to the content that was not previously learned from the image based on their respective areas of interest. This allows the feature information of different categories and the differences in different regions to interact, communicate, and collaborate through the cross-attention mechanism, aggregating the personnel feature information in the image. Therefore, in each layer of the decoder, they can communicate with each other and then repeat this pro-

cess in the next layer until a consensus is reached on the region where the features of port-like personnel appear. It is these deviations in the object queries that allow each object query to focus on the content of its target area, enabling the model to better perceive and capture the global position feature information of port-like personnel. The object queries are transformed into output embeddings through the decoder and then passed through a feed-forward network. They are independently decoded into corresponding box coordinates and class labels, resulting in N final prediction boxes (where N = 100).

Stage 3: Ensemble prediction matching

The output embeddings from the decoder are input into a feed-forward network (FFN). On the one hand, this leads to in the prediction of N normalized bounding boxes of fixed size (including center coordinates, height, and width). On the other hand, class labels are predicted using a Softmax function. The FFN consists of three parts: (1) a ReLU function; (2) a 3-layer MLP with a d-dimensional hidden layer and (3) a linear layer. The process involves matching the set of N predicted Bbox (e.g., 100) output from the feed-forward network with the ground truth Bbox. This is done in a way that minimizes the cost between the predicted Bbox and the ground truth Bbox, thereby transforming it into an optimal bipartite graph matching problem. The calculation formulas are shown in Eq. (6) and Eq. (7). N is typically set to be larger than the typical number of objects in the image. Therefore, an additional label ϕ is used to represent the detection of non-ROI areas (i.e., background). After all the predicted personnel Bbox and all the ground truth Bbox have been traversed, that is, when the predicted personnel Bbox and the ground truth Bbox have a unique match, this approach can avoid the use of non-maximum suppression (NMS) to remove duplicate detection boxes. Finally, a bipartite graph matching loss is calculated for all matched Bbox.

$$\hat{\eta} = \underset{\eta \epsilon \sigma_N}{\arg min} \sum_{i}^{N} \varphi_{mh}\left(a_i, \hat{a}_{\varphi(i)}\right) \tag{6}$$

$$\varphi_{mh}\left(a_i, \hat{a}_{\varphi(i)}\right) = -1_{\{\mathcal{J}\}}\hat{P}_{\eta}\left(l_i\right) + 1_{\{\mathcal{J}\}}\varphi_{box}\left(d_i, \hat{d}_{\eta(i)}\right) \tag{7}$$

where $\varphi_{mh}\left(a_i, \hat{a}_{\varphi(i)}\right)$ represents a pair-wise matching cost between the calculated real ground personnel location data a_i and the personnel location data $\varphi_{(i)}$. $\eta \epsilon \sigma_N$ represents the range of matching between the predicted personnel Bbox and the real ground personnel Bbox. a_i represents the sequence of real personnel ground data, including ϕ. for the i-th real ground data $a_i=(l_i, d_i)$, where l_i represents the person label. d_i represents the center coordinates, width and height of the ground truth Bbox expressed as an array of $d_i \in [0, 1]^4$. $\hat{a}_{\varphi(i)} = \{\hat{a}_i\}_{i=1}^{N}$ represents the sequence of Bbox predicted by the network model as N. $\hat{a}_{\varphi(i)} = \left(\hat{l}_i, \hat{d}_i\right)$ represents the i-th predicted Bbox size and label, where $\hat{P}_{\eta}(l_i)$ represents the probability of belonging to class l_i, and $\hat{d}_{\eta}(i)$ represents the size of the prediction Bbox. φ_{box} represents the loss between the predicted Bbox and the detection Bbox. The Hungarian loss is calculated to obtain scores, categories, center coordinates, width and height of the predicted Bbox while the prediction Bbox is matched with ground truth counterpart. This is essentially a linear combination of the loss of category prediction and Bbox loss. In that way, the Hungarian loss (φ_{Hn}) of all pairs is calculated using Eq. (8):

$$\varphi_{Hn}\left(a_i, \hat{a}\right) = \sum_{i}^{N}\left[-log\,\hat{P}_{\hat{\eta}(i)}(l_i) + 1_{\{\mathcal{J}\}}\varphi_{box}(d_i, \hat{d}_{\hat{\eta}}(i))\right] \tag{8}$$

where $\hat{\eta}$ represents the optimal box match, and $\hat{P}_{\hat{\eta}(i)}(l_i)$ represents matching cost probability. The $\varphi_{box}(d_i, \hat{d}_{\hat{\eta}}(i))$ represents the Intersection over Union (IoU) loss, and log-probability is reduced into 10 times lower when the l_i equals ϕ. We obtain the final detected personnel imaging position from candidate positions by setting a confidence threshold.

C. Trajectory Association and Extraction

Stage 1: Track handling and state estimation

The personnel position in consecutive frames can be obtained (i.e., the personnel detection Bbox) with the help of personnel detector. We propose a port personnel tracker to perform data association and matching procedure considering that the detection results of each image are not linked with each other. The improved DeepSort is introduced based on port personnel movement status with the help of Kalman filter. The model is initialized with personnel detection results in the current frame. The personnel position in the port images is formulated with Eq (9).

$$L = \left(x, y, \gamma, \xi, \dot{x}, \dot{y}, \dot{\gamma}, \dot{\xi}\right) \tag{9}$$

where x and y represent the center coordinates of the port personnel position, aspect ratio γ, ξ is the height of the port personnel image, $(\dot{x}, \dot{y}, \dot{\gamma}, \dot{\xi})$ represents the velocity corresponding to each coordinate direction; the Kalman filter estimates the current position information based on the historical position information, so only the spatial position information is used.

Stage 2: Data association

In order to perform personnel location data association, we use the weighted fusion results of the fused motion model and appearance feature information. The motion model predicts the speed and position information of the next frame based on previous personnel detection, which allows better motion state differentiation for varied people. The motion model uses Bayesian probability theory to combine prior knowledge with real-time observation data for state estimation. This is achieved by predicting the velocity, acceleration, and position information of a person in the next frame based on the previous person detection, which enhances the distinction between individual motion states. Besides, the similarity of appearance feature information is used to avoid ID change for the same person. The personnel trajectory data is further rectified using the IOU data match mechanism.

Moreover, we employ a memory mechanism to correct the trajectory loss outlier when the personnel match failure error occurs twice. We use two metrics, the Mahalanobis distance and the cosine distance, to achieve the correlation between the person motion state and the external feature information. More specifically, the matching between the predicted person and the

newly detected person is accomplished by the Kalman filter, and thus enables the association of data from the same person. To ensure the stability and accurate correlation of motion states, we use the Kalman filter to predict personnel imaging positions and speed information for the following frame based on previous detection results. Therefore, we use Mahalanobis distance to calculate similarity of motion states, which can further measure match level between the predicted and current port personnel detected position. In that manner, we can achieve minimum deviation between the predicted bounding box (i.e., predicted position) and the detected bounding box (i.e., detected personnel position). The formula for calculating the Mahalanobis distance is shown in Eq. (10):

$$g^{mahdis}(i,j) = \left(g_j - p_i\right)^T Q_i^{-1} \left(g_j - p_i\right) \tag{10}$$

where $g^{mahdis}(i,j)$ represents the Mahalanobis distance between the detected value of the port personnel and the predicted value, g_j represents the detected position of the port personnel, p_i represents the predicted position of the port personnel, Q_i is the covariance matrix between detection Bbox and track Bbox.

The Kalman filter can successfully estimate object motion position while object velocity is a constant value. The distance between the tracker predicted box and the detected box is large when the object movement status shows obvious variation (i.e., sudden acceleration or deceleration). The Mahalanobis distance may fail to accurately measure the error. To address the problem, we design a deep appearance feature matching method (i.e., CNN model) to extract the appearance features of port personnel in the images. It is found that appearance features for same person are quite similar, and vice versa. The cosine distance is used to calculate similarity in the study. It is worth noting that the smaller cosine distance indicates that the appearance features between the pedestrians detected in the previous frame and the current frame are closer. The calculation formula for the cosine distance is shown in Eq. (11):

$$g^{cosdis}(i,j) = \min\left(1 - e_j^T e_k^{(i)} \mid e_k^{(i)} \in R_{ij}\right) \tag{11}$$

where $g^{cosdis}(i,j)$ represents the minimum cosine distance between the position detection value and the predicted value of the port personnel. $e_k^{(i)}$ represents the feature vector predicted by the i-th port personnel preserving the features that have been successfully tracked for k times. e_j^T represents the transpose of the image feature vector of port personnel detected position. R_{ij} represents the set of image feature vectors of all tracked port personnel.

After the movement state of the person has been correlated by filtering the Mahalanobis distance, the appearance features are also matched and correlated by cosine distance to obtain the minimum appearance feature difference. The two indicators mentioned above are weighted together to achieve the optimal one-to-one match. The calculation formula for the weighted integration of Mahalanobis distance and cosine distance is shown in Eq. (12):

$$g(i,j) = \ell * g^{mahdis}(i,j) + (1-l)\, g^{cosdis} \quad (i,j) \tag{12}$$

where ℓ represents the weight coefficient, which can be used to adjust the weight of the distance between the appearance feature and motion states, and $g(i,j)$ represents the cascade matching distance after weighting adjustment. Finally, the first data association is completed by cascading the detected position and the predicted position of the port personnel, and the motion trajectory across the image frames is preliminarily obtained.

For the unmatched personnel position detection value and port personnel position prediction value, the Hungarian algorithm based on Intersection over Union (IoU) is used for secondary matching. The calculation formula for the IoU is shown in Eq. (13):

$$IoU = \frac{area\,(pd \cap dec)}{area\,(pd \cup dec)} \tag{13}$$

where dec represents the detection Bbox area of each frame for port personal by the port personnel detector. pd represents the predicted Bbox area for each frame of port personnel, thereby completing the second data association matching.

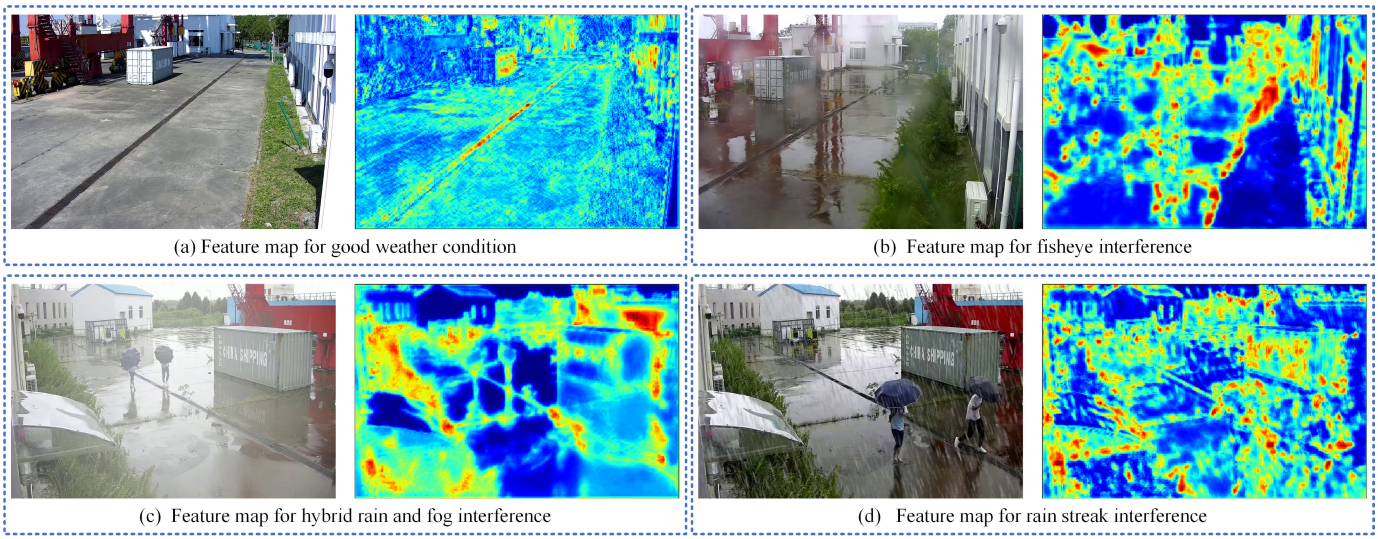
Port personnel data samples may be unsuccessfully linked due to person tracking loss outlier, new trajectory segmentation generation, trajectory deletion, etc. To solve this problem, we have set a memory storage module in DeepSort to save this occluded and distorted stable trajectory information, which includes the ID and feature information of each personnel. The stored information will undergo additional data association matching in the next frame. The port personnel trajectory data will be updated when the data association match procedure is conducted, and thus we can obtain holistic yet accurate trajectory for each personnel.

IV. Experimental Design

A. Data Description

For the purpose of model generalization, we incorporate the COCO 2017 dataset into model training procedure, which includes 118000 training images and 5000 validation images [33]. Each image in the dataset is labeled with multiple object instances, which involves over 80 different object categories (including pedestrians, boats, cars, and other categories). Besides, we collected three port-like videos in our university with different visibility conditions (including raindrop, rain streak, hybrid interference of rain and fog). The image resolution for each collected video is 1280×720, and the frame rate is 25 frames per second (fps). Each scenario involves challenges such as small target personnel detection, personnel image distortion and occlusion, low visibility, etc. Video #1 mainly involves fisheye lens interference, while Video #2 is obtained under low visibility conditions with hybrid interferences from both rain and fog. The video #3 was collected to test model performance under rain streak condition.

To evaluate our model performance against adverse weather, we have collected video #4 in the night port-like scenario and the data sample number was approximately ten-folds larger than those of the three videos. Moreover, people in the video also randomly walked while different people often occluded with each other in the image sequences. The video #5 was



(a) Feature map for good weather condition

(b) Feature map for fisheye interference

(c) Feature map for hybrid rain and fog interference

(d) Feature map for rain streak interference

Fig. 3. Feature maps for port images under different weather conditions.

TABLE I

INFORMATION OF PORT SURVEILLANCE VIDEOS

Information	video #1	video #2	video #3	video #4	video #5
interference scene	raindrop	hybrid of rain and fog	rain streak	night	normal
data sample	1442	1112	1462	15111	1256
frame rate (fps)	25	25	25	25	25
resolution	1280×720	1280×720	1280×720	1280×720	1280×720
small target detection	√	√	√	√	/
image distortion	√	√	√	/	/
visibility	√	√	√	√	/

(symbol √ indicates the situation that exists in the video)

collected to further testify model performance under good weather condition, and trajectory data samples were 1256. Both framerate and image resolution for the video #5 were same to those of video #1. More details for each collected video can be found in Table I.

Details for the three videos can be found in Table I. Note that each video is also involved with image distortion and small target (i.e., personnel imaging size is small) interferences. The proposed method is implemented with PyTorch 1.4.0 framework and Python 3.7. The operating system is Ubuntu 20.04 OS, and the CPU is Intel(R) Xeon(R) Gold 6230R CPU @ 2.10GHz. The GPU used for the experimental platform is Quadro RTX 5000. Figure 3 demonstrates feature maps for the collected videos under varied interferences. Figure 3(a) shows the port environmental feature map under good weather conditions, while the figure 3(b), (c) and (d) show the feature maps of raindrop, hybrid interference of rain and fog, rain streak weather condition.

B. Evaluation Indicators

We manually obtain ground truth port personnel imaging trajectories from three videos for the purpose of model performance comparison. Each person in a video clip is assigned with a unique ID for the purpose of trajectory extraction. We employ six indicators to verify model performance, which include multiple-object tracking accuracy (MOTA), identification of the same personnel ID in each Bbox (IDF_1), value of judging the same target identification recall of personnel ID in each Bbox frame (IDR), identification precision of personnel ID in each Bbox frame (IDP), ID switch (ID_SW) and fps. The MOTA represents the model tracking accuracy; and ID_SW demonstrates port personnel tracking ID variation frequency. The formulas for calculating MOTA and IDF_1 are shown in Eq. (14) and Eq. (15), respectively. The IDP represents the identification accuracy (see Eq. (16)) and IDR represents the identification recall rate of the ID for the personnel in the port images (see Eq. (17)). Fps represents model computational each frame average time consumption (see Eq. (18)). It is considered that the extracted port personnel trajectory is closer to the ground truth counterpart with larger MOTA, IDF_1, IDR, IDP, fps and smaller ID_SW.

$$MOTA = 1 - \frac{fn + fp + \omega}{t} \tag{14}$$

$$IDF_1 = \frac{2idtp}{2idtp + idfp + idfn} \tag{15}$$

$$IDP = \frac{idtp}{idtp + idfp} \tag{16}$$

$$IDR = \frac{idtp}{idtp + idfn} \tag{17}$$

$$fps = \frac{\sum_{n=1}^{fr} \frac{1}{CT_{fr}}}{fr} \tag{18}$$

where fn represents miss-tracking ID number, and fp represents ID wrongly-tracking number. The ω represents the port personnel ID switch number, and t represents ground truth ID for the port personnel. The $idtp$ ($idfp$) indicates the number that predicted port personnel ID well-matches (wrongly-matches) with ground truth ID. The $idfn$ represents the number of times the tracking algorithm failed to assign the correct ID. The CT_{fr} demonstrates time cost for each individual frame, and fr is frame number for each collected port video.

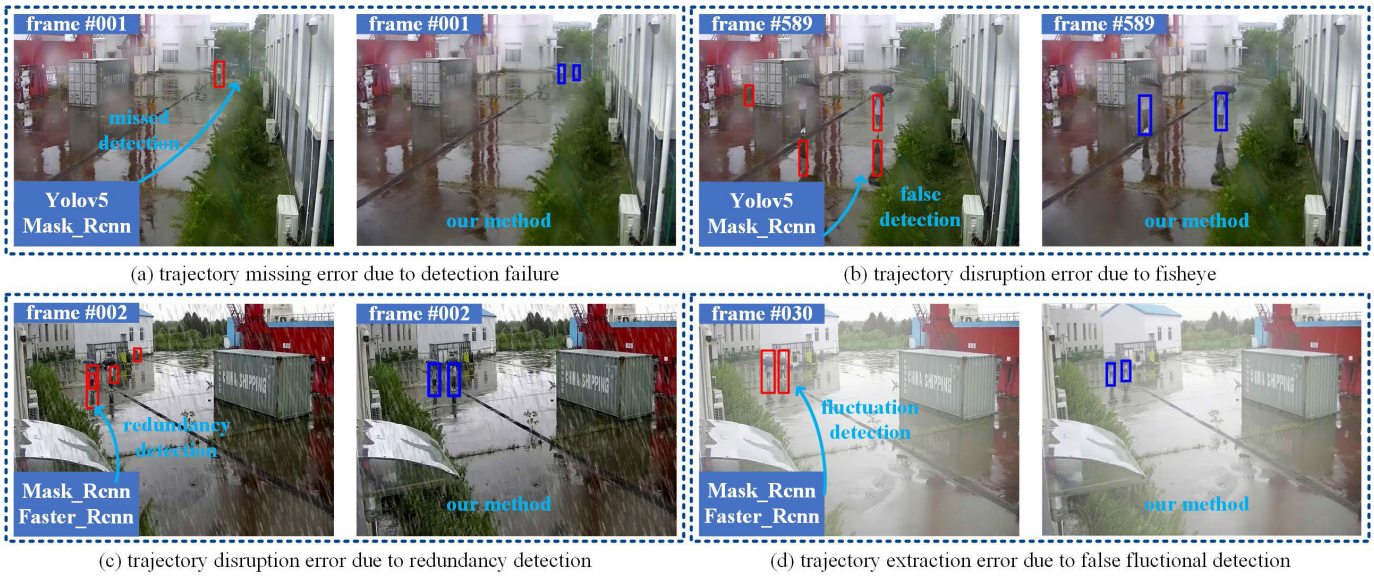


Fig. 4. Typical port personnel detection errors for different models under different rainy interferences.

V. Experimental Results

A. Experimental Results About Port Personnel Detection

Figure 4(a) indicated that miss-detection can be easily found in port personnel detection under interferences of image distortion, small object imaging size, etc. As shown in figure 4(b) and (c), low visibility interference triggers false detection (i.e., non-personnel target was wrongly detected as people) and redundancy detection (one person was assigned with multiple Bbox). The port personnel pixels are very close to backgrounds under complex port area environment interference (i.e., low visibility of image features and details) as shown in figure 4(d).

Both of figure 4(a) and 4(b) indicated that Yolov5 and Mask_rcnn models experienced miss and false detection under image distortion, fisheye and small target interferes. The main reason is that people imaging resolution in the video is too low and personnel features can be hardly identified. Rainy interference causes feature confusion for people and objects in the collected port videos, and thus the detector may fail to identify people from the port videos. The Mask_Rcnn model failed to accurately obtain effective candidate boxes while the target port personnel was occluded by rain due to the fisheye effect. Moreover, the Mask_Rcnn model failed to infer global contextual feature information from local image features, which made the model difficult to recognize port personnel against the personnel reflection appeared in the ground water.

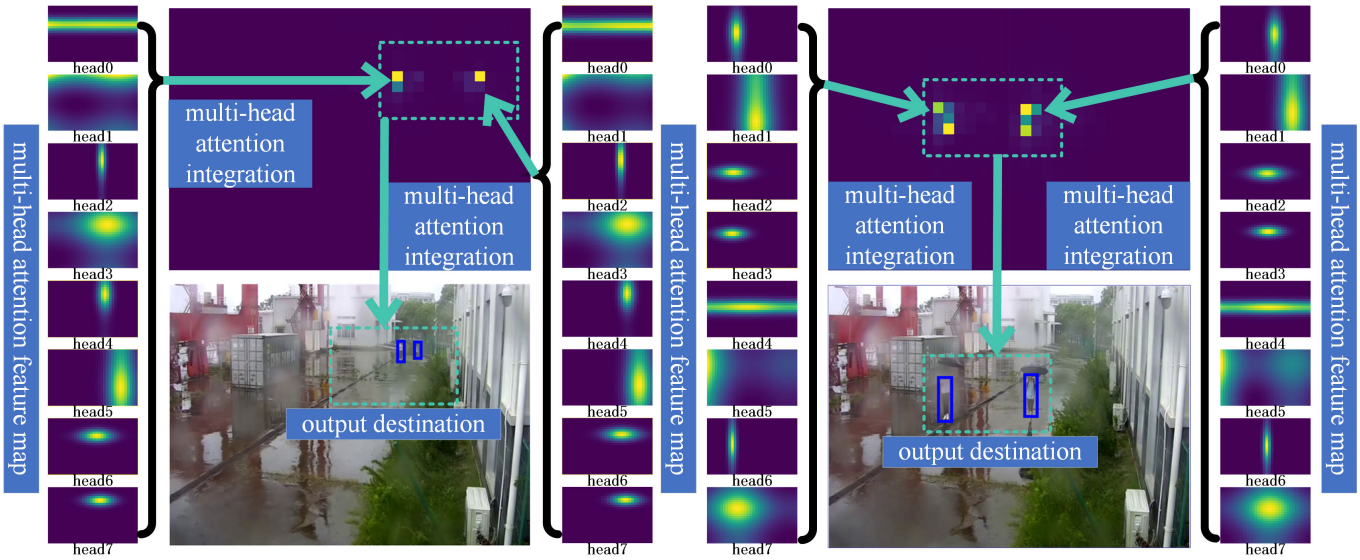
The Yolov5 model was also misled by the port personnel projections in the ground water under rainy weather conditions. In frame #2 of the video #3 (see figure 4(c)), one person was wrongly detected into two Bbox under rainy weather interferences by the Mask_Rcnn and Fast_Rcnn models. It can be infereed that the Mask_Rcnn and Fast_Rcnn models failed to exploit fundamental feature differences among people and objects in the port videos. Figure 4(d) indicated that both of the Mask_Rcnn and Fast_Rcnn experienced Bbox fluctuations due to difficulty in identifying fine-grain features in the hybrid interference of rain and fog and rain streaks. Thus, the detected Bbox (i.e., detected port personnel positions) may fail to well-match with ground truth position.

The object query vector in our proposed farmwork can effectively capture the feature differences among different objects in the port images. The 8-head attention mechanism of the transformer network structure enhanced our framework capabilities of global and local contextual reasoning (see figure 5). In that way, the multi-head attention mechanism can select and set appropriate weight for each head to obtain optimal personnel position identification performance. More specifically, the 8-head attention mechanism helps the model match each predicted Bbox with detected personnel Bbox in one-to-one manner. In that way, trajectory loss and redundance related outliers can be efficiently suppressed by our framework.

B. Experimental Results About Port Personnel Trajectory Extraction

We have further evaluated model performance by extracting port personnel trajectories on the three collected port videos. Our proposed framework was abbreviated as TDM for the purpose of better readability. For the purpose of model performance comparison, we extracted trajectories utilizing Yolov5+DeepSort (YD), Faster_Rcnn+ByteTrack (FRB) and Mask_Rcnn+DeepSort (MRD) to further evaluate model performance. It is worth noting that we visualize the extracted trajectories through point object tracking manner. Note that each personnel in the port video moved with different status, whilst group target tracking based models cannot be introduced straightforwardly to fulfill multiple people trajectory tracking task.

Overall, port personnel trajectories obtained by various models were close to the ground truth counterparts (as shown in figure 6). The yellow (GT 1) and blue (GT 2) bounding boxes demonstrate ground truth personnel trajectories.



(a) position feature maps for personnel with small imaging size (b) position feature maps for personnel with fisheye distortion interference

Fig. 5. Multi-head attention mechanism obtained personnel feature maps for low-visibility port images.

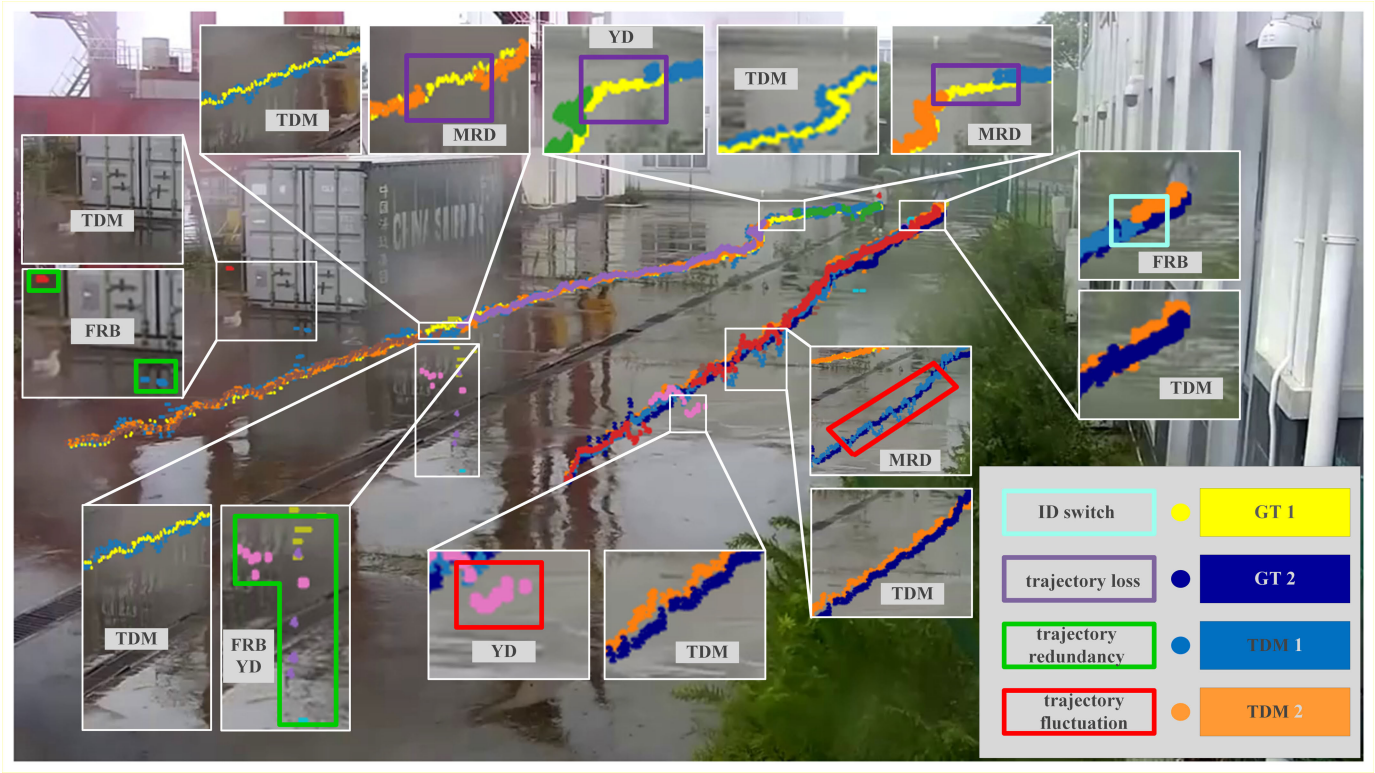
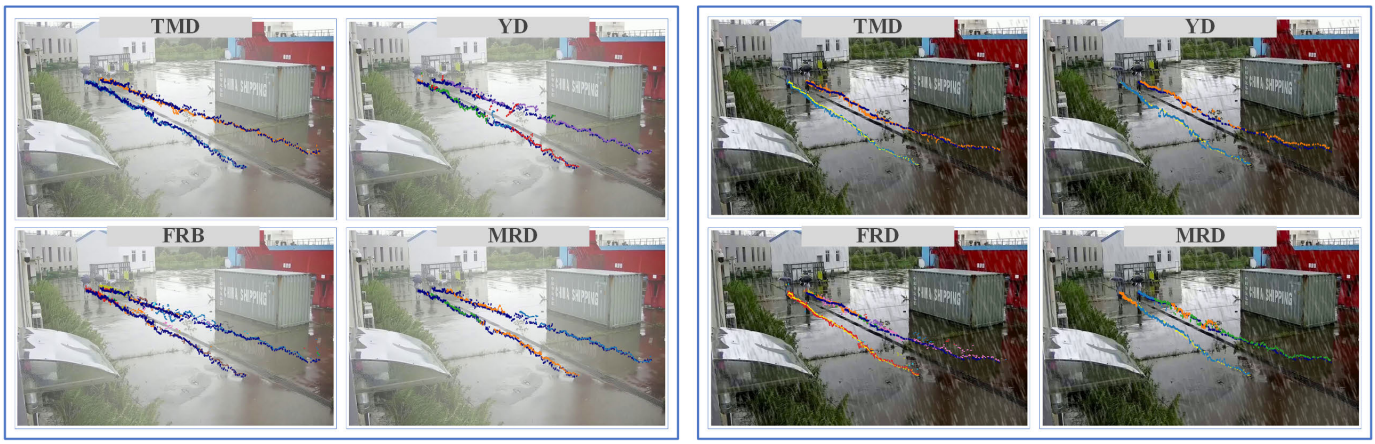


Fig. 6. Port personnel trajectory extraction performance comparison for video #1.

Outliers can be easily observed when zooming out personnel trajectory details. Note that port personnel trajectory experienced outliers (e.g., trajectory association error in neighboring images) while trajectory color changed for same people. For instance, personnel trajectory may be missing for a long time due to raindrop and video imaging distortion interferences. Personnel imaging occlusion may lead to ID switch outlier.

The trajectory data obtained by our proposed TDM model was quite close to those of the ground truth counterparts. The trajectory data outliers (such as trajectory loss, trajectory fluctuation trajectory redundancy and ID switch) were successfully corrected by our model. The purple box in the figure represents the abnormal trajectory points (i.e., trajectory loss) due to the fisheye effect caused by raindrop. The green box represents trajectory redundancy due to object feature miss (e.g., object occlusion) in the collected port images. The red box and light-cyan box represent trajectory fluctuation and ID switch problem. Meanwhile, the trajectories extracted by different



(a) port personnel trajectory extraction for video #2

(b) port personnel trajectory extraction for video #3

Fig. 7. Port personnel trajectory extraction performance comparison for video #2 and #3.

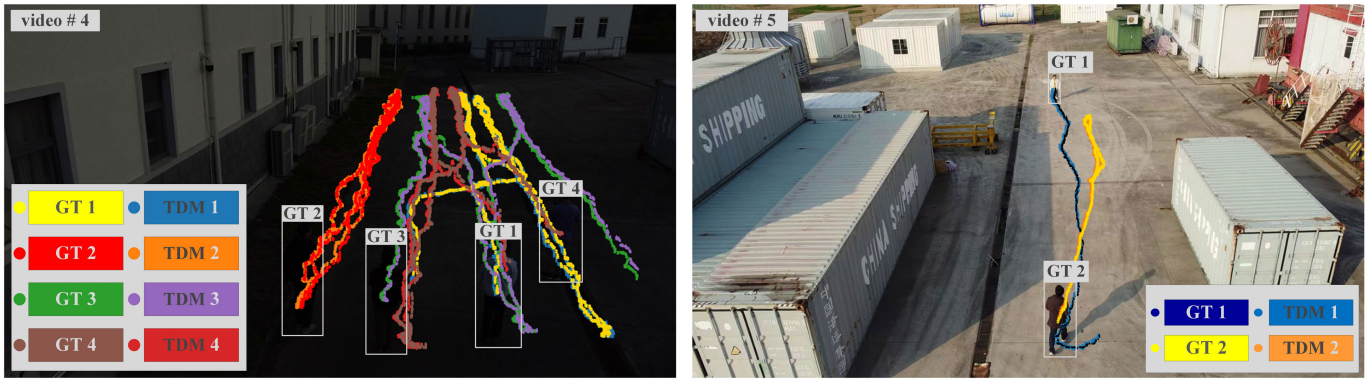
TABLE II
PERFORMANCE STATISTICS OF HUMAN TRAJECTORY EXTRACTION IN PORT ENVIRONMENT UNDER DIFFERENT VISIBILITY

Data	Model	Evaluation indicators					
		MOTA	IDF_1	IDR	IDP	ID_SW	fps
video#1	TDM	**91.47%**	**95.68%**	**94.38%**	**97.01%**	**0**	11.95
	YD	76.49%	69.91%	63.73%	77.42%	6	21.73
	FRB	79.26%	79.14%	81.97%	76.50%	8	6.27
	MRD	83.43%	67.39%	62.41%	73.23%	2	5.78
video#2	TDM	**86.60%**	**93.22%**	**92.18%**	**94.30%**	**0**	11.86
	YD	53.42%	51.54%	45.23%	59.88%	18	23.19
	FRB	71.13%	43.91%	42.25%	45.47%	5	5.93
	MRD	45.68%	32.96%	25.45%	46.78%	3	5.57
video#3	TDM	**99.04%**	**99.52%**	**99.38%**	**99.66%**	**0**	11.92
	YD	78.80%	88.21%	79.34%	99.32%	0	22.39
	FRB	77.50%	70.21%	75.85%	65.35%	38	5.04
	MRD	87.28%	55.20%	54.10%	56.34%	16	5.19
video#4	**TDM**	**99.44%**	**99.72%**	**99.60%**	**99.84%**	**0**	10.56
	YD	76.05%	54.52%	48.03%	63.04%	13	24.30
	FRB	87.86%	73.89%	73.02%	74.79%	4	4.63
	MRD	78.45 %	77.67 %	83.52 %	72.59 %	11	4.34
video#5	TDM	99.52%	99.76%	99.52%	100%	0	12.82
	YD	99.20%	99.60%	99.20%	100%	0	28.16
	FRB	98.70%	99.08%	99.25%	98.91%	0	8.52
	MRD	94.24%	97.17%	98.68%	95.70%	0	7.89

models are represented in a gray bottom box in figure 6. However, the TDM framework proposed in the study did not experience abnormal trajectory points or data loss errors.

Figure 7 demonstrated trajectory extraction performance for video #2 and #3 for different models. It can be observed that varied models showed similar performance compared to those in video #1. Table II demonstrated trajectory extraction performance for three videos in a quantitative manner. The MOTA indicator for our proposed TDM model was 91.47% for video #1, which was approximately 10% higher than those of the YD, FRB and MRD models. The indicators of IDF_1, IDR IDP and ID_SW, obtained by our proposed model, for video #1 were 95.68%, 94.38%, 97.01% and 0. It can be found that proposed model outperformed the counterparts for video #1 in terms of MOTA, IDF_1. The average fps for the YD model was 22.43, which was approximately two times larger than that of the TDM model. Because YOLOv5 is a one-stage detector with fast detection speed. However, its lower detection accuracy may result in unassociated personnel data, further reducing the time and cost of model inference.

We have further testified our proposed TDM model performance on the port-like video captured in the evening. The visibility condition in video #4 was low, and four people walked on-site. The ground truth trajectory for each person was labeled GT1, GT2, GT3, GT4 as shown in figure 8(a). The trajectory distributions for the ground truth data demonstrated that the person #1 walked back and forth. In that way, trajectory loss was easily triggered (i.e., ID switch), which significantly challenged model robustness. We evaluated port personnel trajectory extraction performance for different



(a) port personnel trajectory extraction for video #4

(b) port personnel trajectory extraction for video #5

Fig. 8. Trajectory extraction performance for different models of videos #4 and #5.

TABLE III
ABLATION EXPERIMENT RESULTS WITH DEEP ASSOCIATION METRIC MODULE AND MEMORY MODULE

Data	Model	Evaluation indicators					
		MOTA	IDF_1	IDR	IDP	ID_SW	fps
video#1	TS	61.03%	46.48%	52.08%	41.96%	14	13.42
	TSD	84.47%	67.30%	68.10%	66.53%	8	12.75

models with MOTA, IDR, IDF_1, IDP and ID_SW indicators (as shown in table II).

The previous four metrics of our proposed TDM model in video #4 were 99.44%, 99.60%, 99.72% and 99.84% respectively, which are 11.58%, 25.83%, 26.58% and 25.05% larger than those of the FRB mode. Moreover, our proposed model suppressed the ID switch weakness due to that ID_SW value was 0. In comparison, the remaining YD, FRB, and MRD models for the ID_SW indicator were 13, 4, and 11, respectively. From the perspective of time consumption, the YD model is a one stage trajectory extraction framework, and the frame rate of the YD model is approximately 22.39. Our collected port-like videos were public-accessible on the website *https://github.com/xinqiangchentraffic/TDM_demo_video*.

From perspective of five statistical indicators, our proposed model obtained better performance compared to those of video #1, #2, #3 and #4. Table II indicated that MOTA, IDF_1, IDR, IDP and ID_SW obtained by the proposed model (for video #5) were quite close to 100%. Figure 8(b) showed the ground truth and TDM model obtained trajectories, which have confirmed that our proposed framework can obtain satisfactory performance under good visibility condition. Thus, it can be concluded that our proposed framework can successfully extract port personnel trajectory under both adverse and good visibility weather conditions.

We considered that our model obtained better performance compared to the YD model because the accuracy of TDM model trajectory extraction was higher. The time cost for the TDM model was 11.91, which was about two-fold larger than those of the FRB and MRD modes. Statistical indicator distributions for both of video #2 and #3 showed similar variation tendency in comparison with those of video #1.

It is noted that statistics (i.e., MOTA, IDF_1, IDR, IDP, ID_SW) for video #2 are slightly lower than those of video #1 and #3. The main reason was that image whitening phenomenon was more obvious under rain and fog interferences. Port personnel visibility and the corresponding features were potentially polluted or corrupted in the image sequences and thus object imaging boundaries were interweaved with background related area. In sum, it can be safely concluded that our proposed model obtained satisfied performance for fulfilling port personnel trajectory extraction task under varied rainy interferences.

C. Ablation Experiments

We have implemented additional ablation experiment on the proposed TDM model in video #1 to further verify our model performance. The first ablation experiment was conducted by adding motion state estimation module. The second ablation experiment was implemented by integrating deep association metric module. Table III indicated that motion state estimation (MES) and deep association metric module (DAM) can lead to model performance. With the help of deep association metric module, personal trajectory can be successfully associated with neighboring frames while intrinsic personal imaging features can be exploited. The statistical indicators suggested that the DETR+MES+DAM (TSD) model outperformed the DETR+MSE (TS) model with 15% performance improvement. Moreover, the ID switch phenomenon was successfully tackled by the TSD model. In comparison, the MOTP and IDF_1 obtained by our proposed TDM model were both larger than 90% (which can be refereed to table II TDM). The ablation results also verified our proposed model performance.

VI. Conclusion

Port surveillance videos provide informative spatiotemporal data for supporting port management efficiency. Personnel imaging trajectory extraction is of great practical significance for ensuring port safety and security. Adverse weather conditions (e.g., rain, fog) challenge spatial-temporal data (such as personnel trajectory) extraction accuracy from port surveillance videos. The study proposed an ensemble transformer and memory-improved DeepSort deep learning model to extract port personnel imaging trajectories under varied rainy interferences. The proposed framework collected global features from the input port image sequences using the encoder-decoder module of the transformer structure. In this way, we can further obtain port personnel information (such as ID and positions) via the object query and decoder module in the transformer structure. The outliers of extracted port personnel positions were then further corrected via the memory module of the DeepSort tracker in the proposed framework. We verified proposed framework performance in three typical rainy scene videos (raindrop, rain streak, and hybrid weather condition of rain and fog) captured from port-like environments. The aggregated statistics of MOTA, IDF_1, IDR, IDP, ID_SW and fps were 92.37%, 96.14%, 95.31%, 96.99%, 0 and 11.91, which suggested that the proposed framework obtained satisfied performance.

The following directions can be expanded to further enhance model applicability in future. First, port personnel movement status in three videos were homomorphic, and we can further verify model performance under additional personnel motion patterns. Second, the people density in the collected port-like videos were not large, and port scenario verification under large density scenario deserves our further attentions. Last but not least, we further evaluate model performance under maritime environments (e.g., ship trajectory extraction under adverse weather interferences).

References

[1] Z. H. Munim, O. Duru, and A. K. Y. Ng, "Transhipment port's competitiveness forecasting using analytic network process modelling," *Transp. Policy*, vol. 124, pp. 70–82, Aug. 2022.

[2] Ç. Iris and J. S. L. Lam, "A review of energy efficiency in ports: Operational strategies, technologies and energy management systems," *Renew. Sustain. Energy Rev.*, vol. 112, pp. 170–182, Sep. 2019.

[3] X. Chen, S. Liu, R. W. Liu, H. Wu, B. Han, and J. Zhao, "Quantifying Arctic oil spilling event risk by integrating an analytic network process and a fuzzy comprehensive evaluation model," *Ocean Coastal Manage.*, vol. 228, Sep. 2022, Art. no. 106326.

[4] C. H. Bahnsen and T. B. Moeslund, "Rain removal in traffic surveillance: Does it matter?" *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2802–2819, Aug. 2019.

[5] M. Li, X. Cao, Q. Zhao, L. Zhang, and D. Meng, "Online rain/snow removal from surveillance videos," *IEEE Trans. Image Process.*, vol. 30, pp. 2029–2044, 2021.

[6] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2482–2491.

[7] A. Palvanov and Y. Im Cho, "DHCNN for visibility estimation in foggy weather conditions," in *Proc. Joint 10th Int. Conf. Soft Comput. Intell. Syst. (SCIS) 19th Int. Symp. Adv. Intell. Syst. (ISIS)*, Dec. 2018, pp. 240–243.

[8] S. Zang, M. Ding, D. Smith, P. Tyler, T. Rakotoarivelo, and M. A. Kaafar, "The impact of adverse weather conditions on autonomous vehicles: How rain, snow, fog, and hail affect the performance of a self-driving car," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, pp. 103–111, Jun. 2019.

[9] W. Yang et al., "Advancing image understanding in poor visibility environments: A collective benchmark study," *IEEE Trans. Image Process.*, vol. 29, pp. 5737–5752, 2020.

[10] X. Chen, Z. Wang, Q. Hua, W.-L. Shang, Q. Luo, and K. Yu, "AI-empowered speed extraction via port-like videos for vehicular trajectory analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 4541–4552, Apr. 2023.

[11] R. S. D. Sousa, A. Boukerche, and A. A. F. Loureiro, "Vehicle trajectory similarity: Models, methods, and applications," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–32, Sep. 2021.

[12] 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.

[13] J. Si, B. Song, J. Wu, W. Lin, W. Huang, and S. Chen, "Maritime ship detection method for satellite images based on multiscale feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6642–6655, 2023.

[14] K. Sirohi, S. Marvi, D. Büscher, and W. Burgard, "Uncertainty-aware panoptic segmentation," *IEEE Robot. Autom. Lett.*, vol. 8, no. 5, pp. 2629–2636, May 2023.

[15] Y. Li and J. Košecká, "Uncertainty aware proposal segmentation for unknown object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 241–250.

[16] J. Luo, C. Zhang, and C. Wang, "Indoor multi-floor 3D target tracking based on the multi-sensor fusion," *IEEE Access*, vol. 8, pp. 36836–36846, 2020.

[17] S. Plangi, A. Hadachi, A. Lind, and A. Bensrhair, "Real-time vehicles tracking based on mobile multi-sensor fusion," *IEEE Sensors J.*, vol. 18, no. 24, pp. 10077–10084, Dec. 2018.

[18] J. Bai, S. Li, L. Huang, and H. Chen, "Robust detection and tracking method for moving object based on radar and camera data fusion," *IEEE Sensors J.*, vol. 21, no. 9, pp. 10761–10774, May 2021.

[19] Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, and H. Wang, "Learning deep multi-level similarity for thermal infrared object tracking," *IEEE Trans. Multimedia*, vol. 23, pp. 2114–2126, 2021.

[20] S. Panicker, A. K. Gostar, A. Bab-Hadiashar, and R. Hoseinnezhad, "Tracking of targets of interest using labeled multi-Bernoulli filter with multi-sensor control," *Signal Process.*, vol. 171, Jun. 2020, Art. no. 107451.

[21] H. Wu et al., "Contrastive learning for compact single image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10546–10555.

[22] M. A. Kenk, M. Hassaballah, M. A. Hameed, and S. Bekhet, "Visibility enhancer: Adaptable for distorted traffic scenes by dusty weather," in *Proc. 2nd Novel Intell. Lead. Emerg. Sci. Conf. (NILES)*, Oct. 2020, pp. 213–218.

[23] M. Hassaballah, M. A. Kenk, K. Muhammad, and S. Minaee, "Vehicle detection and tracking in adverse weather using a deep learning framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4230–4242, Jul. 2021.

[24] R. Quan, X. Yu, Y. Liang, and Y. Yang, "Removing raindrops and rain streaks in one go," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9143–9152.

[25] Z. Tu et al., "MAXIM: Multi-axis MLP for image processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5759–5770.

[26] E. Benli, R. L. Spidalieri, and Y. Motai, "Thermal multisensor fusion for collaborative robotics," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 3784–3795, Jul. 2019.

[27] M. Krišto, M. Ivasic-Kos, and M. Pobar, "Thermal object detection in difficult weather conditions using YOLO," *IEEE Access*, vol. 8, pp. 125459–125476, 2020.

[28] M. P. Muresan, S. Nedevschi, and R. Danescu, "Robust data association using fusion of data-driven and engineered features for real-time pedestrian tracking in thermal images," *Sensors*, vol. 21, no. 23, p. 8005, Nov. 2021.

[29] D. Yuan, X. Shu, Q. Liu, and Z. He, "Aligned spatial–temporal memory network for thermal infrared target tracking," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 70, no. 3, pp. 1224–1228, Mar. 2023.

[30] B. Iepure and A. W. Morales, "A novel tracking algorithm using thermal and optical cameras fused with mmWave radar sensor data," *IEEE Trans. Consum. Electron.*, vol. 67, no. 4, pp. 372–382, Nov. 2021.
[31] Z. Ouyang, J. Cui, X. Dong, Y. Li, and J. Niu, "SaccadeFork: A lightweight multi-sensor fusion-based target detector," *Inf. Fusion*, vol. 77, pp. 172–183, Jan. 2022.
[32] J. Shao, B. Du, C. Wu, M. Gong, and T. Liu, "HRSiam: High-resolution Siamese network, towards space-borne satellite video tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 3056–3068, 2021.
[33] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer 2014, pp. 740–755.

Xinqiang Chen (Member, IEEE) received the Ph.D. degree from Shanghai Maritime University, China, in 2018. From September 2015 to September 2016, he was a Visiting Student with the Smart Transportation Applications and Research Laboratory, University of Washington, Seattle, WA, USA. His research interests include transportation image processing and smart ship and maritime traffic situation awareness.

Chenxin Wei is currently pursuing the M.S. degree with the Institute of Logistics Science and Engineering, Shanghai Maritime University, China. His research interests include computer vision, traffic image processing, and intelligent transportation systems.

Yang Yang is currently a Post-Doctoral Fellow with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His research interests include traffic safety, energy conservation and emission reduction, transportation planning, and cooperation vehicle-infrastructure systems.

Lijuan Luo received the Ph.D. degree from the Beijing University of Posts and Telecommunications, China, in 2015. She is currently an Associate Professor with Shanghai International Studies University, China. Her research interests include machine learning, neurocomputing, and data decision making.

Salvatore Antonio Biancardo received the Ph.D. and Doctor Europaeus degrees in civil engineering from the University of Naples Federico II, Italy, in 2016. He is currently an Assistant Professor with the Department of Civil, Construction and Environmental Engineering, University of Naples Federico II. His research interests include BIM for infrastructures, road pavement materials, and transportation safety.

Xiaojun Mei (Member, IEEE) received the Ph.D. degree from Shanghai Maritime University, Shanghai, China, in 2021. From September 2019 to September 2020, he was a Visiting Ph.D. Student with the Institute for Systems and Robotics, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. From November 2021 to October 2023, he was a Post-Doctoral Fellow with Shanghai Maritime University. He is currently an Associate Professor with the Merchant Marine College, Shanghai Maritime University.