IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 April 2023, accepted 18 May 2023, date of publication 23 May 2023, date of current version 31 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3279314

RESEARCH ARTICLE

Research on Siamese Object Tracking Algorithm Based on Knowledge Distillation in Marine Environment

YIHONG ZHANG, (Senior Member, IEEE), QIN LIN, HUIZHI TANG, AND YINJIAN LI
College of Information Science and Technology, Donghua University, Shanghai 200000, China

Corresponding author: Qin Lin (2211904@mail.dhu.edu.cn)

This work was supported in part by the Fundamental Research Funding for the Central Universities of the Ministry of Education of China under Grant 18D110408, and in part by the National Natural Science Foundation of China (NSFC) under Grant 18K10454.

ABSTRACT Siamese networks have gained considerable attention for object tracking due to their balance of speed and accuracy. However, existing Siamese tracking algorithms have been too rigid in their predictions of bounding box tags and lack uncertainty estimation, resulting in poor tracking performance in marine environments, particularly those with waves. To improve the effectiveness of trackers in marine environments, this study proposes a Siamese distillation network. First, to address the issue that the presence of waves and other disturbances may result in target loss or inaccuracy when tracking the target, the concept of a probability distribution of the bounding box is introduced in this study, which transforms the standard Dirac delta distribution of the bounding box into a probability distribution of the bounding box, effectively reducing the impact of interference on tracking performance and improving target location accuracy. Second, we chose ResNet100 as the backbone network to obtain richer features for localization. Finally, this work offers a knowledge distillation approach to further enhance the tracking accuracy and model performance, while considering the impact of the model's number of parameters and computational amount on tracking performance. This network outperforms most trackers in terms of accuracy, according to extensive experimental results, and performs well on the target tracking benchmark and marine dataset annotated in this study. Specifically, this network achieved the highest accuracy value of 0.612 compared to other Siamese networks, resulting in a 2.5% increase compared to original baseline network. This suggests that the proposed algorithm is practical.

INDEX TERMS Bounding box probability distribution, knowledge distillation, object tracking, siamese network.

I. INTRODUCTION

Water safety has become a more pressing issue due to the rapid development of the maritime transport sector and improvements in people's quality of life. Although safety concerns when working at sea have recently gained more attention, rescue operations for those who have fallen overboard remain crucial [1]. It is widely recognized that search and rescue operations in the sea's complex environments can be challenging. The search for individuals who have fallen into the water cannot be performed exclusively with the naked eye, especially in bottomless seas [2]. Accurate and fast position prediction can have significant supplemental value for rescuers and can maximize the protection of the drowning person's life. The development of rescue Unmanned Aerial Vehicles (UAVs) has partially alleviated the difficulties encountered during rescue operations at sea [3].

Rescuers can utilize UAVs to locate victims who have fallen into the sea, and tracking mostly depended on the tracking algorithms [4]. Visual object tracking has historically been one of the primary applications of computer vision and has been extensively studied [5]. Owing to the

The associate editor coordinating the review of this manuscript and approving it for publication was Charalambos Poullis.

increased demand for tracking algorithms across several industries, multiple industries have higher expectations for tracker speed, accuracy, and robustness.

With the development of deep learning, object-tracking algorithms are constantly improving, and algorithms for maritime-target tracking is also emerging. For the complicated environments of deep learning techniques, tracking speed and scale change. Zhang et al. [6] presented depth fusion of multi-scale related filtering sea target tracking algorithms, whereas existing tracking methods only employ single-layer depth features or manual fusing of multi-layer features. The target can be effectively moved by water using this technique. Wu et al. [7] proposed a scale-adaptive sea surface target tracking algorithm based on deep learning to classify samples based on whether the sample's center point falls inside the actual target box and then directly returns to the distance from the center point to the target box to predict the position and scale of the target box. This algorithm was designed to effectively address the issue of target-scale adaptation. The Siamese network-based offshore target tracking system proposed by Shi et al. [8] satisfies the demands of real-time performance and accuracy in offshore target tracking. Even though the deep-learning-based maritime target tracking algorithms currently in use function well, a person who falls into the water is not only small but also readily mistaken for marine debris and other organisms owing to the complexity of the marine environment. Waves drastically reduce tracking precision and speed. This is because the bounding box tag will have an uncertain area when waves or other disturbances are present, and this ambiguity will cause the target to be positioned incorrectly. The tracking methods currently used do not consider this uncertainty. Furthermore, the most sophisticated Siamese tracks employed ResNet50 as their backbone network. ResNet50 characteristics can help us extract key details from the majority of common images; however, they are inadequate for object tracking and recognition in challenging maritime environments. To address this problem, this paper proposes SiamKD, a new Siamese tracking algorithm. First, considering that the presence of waves and other distractors may affect the positioning of drowning people, which may lead to the loss or inaccuracy of the target when tracking it, inspired by the idea of bounding box probability distribution modeling [9], this study improves the standard bounding box Dirac delta distribution [10] to the bounding box probability distribution to improve the accuracy of the target location, that is, the accuracy of the predicted bounding box. Second, more sophisticated networks were chosen to create feature maps because they can extract richer features with higher accuracy from deeper and wider networks. This paper also introduces the knowledge distillation [11] model compression method, which considers the fact that deeper and broader networks themselves have more extensive parameters and greater computational complexity, and that the improvement of the general distribution of bounding boxes to the probability distribution of bounding boxes will also significantly increase the parameters of the network to slow down the network training speed and affect the model performance. To verify the algorithm's tracking effectiveness in challenging marine situations, this study annotated sea datasets, which consisted of 100 videos. Our system successfully used numerous open and labeled datasets in experiments, demonstrating its efficacy.

In this study, we propose a Siamese object tracking algorithm based on knowledge distillation, which makes the following contributions:

We propose a Siamese tracking algorithm that replaces the typical Dirac delta distribution of the bounding box with the probability distribution of the bounding box. This significantly reduces the effect of interference on tracking performance and increases the accuracy of the target location.

To make the model lighter and more efficient when using a more sophisticated network to extract rich features, we introduce the model compression approach of knowledge distillation.

We label and use a marine dataset of 100 video sequences to test the tracker's performance.

The algorithm achieves excellent tracking results on the datasets marked in this study and demonstrates strong performance on other datasets such as VOT2018, VOT2019, OTB100, and NFS.

The rest of the paper is organized as follows. In Section II, we briefly review three related works on Siamese network-based visual tracking, bounding box probability distribution, and knowledge distillation. Section III describes the proposed SiamKD method. Section IV describes the experimental details and performance analysis. The conclusions are presented in Section V.

II. RELATED WORK

A. VISUAL TRACKING BASED ON THE SIAMESE NETWORK

Owing to their effectiveness and end-to-end learning capacity, trackers based on Siamese networks have recently attracted considerable interest in visual tracking. The Siamese network structure is a type of neural network that consists of two or more subnetworks. It is distinguished by the fact that it shares the weights of two neural networks and takes two images as inputs. Its main goal is to identify a function that can translate an input image into a target space, where the easy distance is close to the ''semantic'' distance of the input space. Specifically, it seeks to identify a set of parameters for which the similarity measure is small for images that belong to the same category and large for images that do not.

SiamFC [12], the Siamese tracker's first work, is primarily composed of the upper and lower branches. The upper branch is the template branch responsible for generating the initial frame image features. The lower branch is the search branch, which is responsible for generating the features of the subsequent frame images, convolving the two feature maps to generate the final feature map, and then calculating the maximum value in the feature map to achieve the final target tracking

and positioning, from which most subsequent Siamese tracking algorithms are derived. For example, CFNet [13] combines the following advantages: the filtering algorithm and SiamFC algorithm to significantly improve tracker performance. DCFNet [14] is a lightweight Siamese symmetric convolutional network that combines the discriminant Kernel Correlation Filter(KCF) algorithm and convolutional features and can be quickly tracked using only two convolutional layers. These results are further studied. DSiam [15] proposed a network model with a dynamic Siamese symmetry structure that can learn the deformation of a target online and effectively suppress noise. SiamRPN [16] adds a Region Proposal Network(RPN) module based on SiamFC. The feature extraction part of the network was the same as that of the SiamFC. After the feature map is created, the network is split into classification and regression branches. The first determines the category, while the second determines the center point coordinates as well as the length and width of the target box. Da-SiamRPN [17] introduces an effective sampling strategy in the offline training stage to enable the network to learn more discriminative features. It also designs a model to identify non-targets during the prediction stage, which avoids incorrect prediction and positioning of the model and improves discrimination. Sa-Siam [18] improved tracking performance by designing a double Siamese network to extract different features. SiamDW [19] further improved tracking performance by introducing a residual block internal clipping unit in a deeper and wider network. Based on SiamRPN, SiamRPN++ [20] uses the ResNet50 network to replace the original AlexNet, adds a multilayer fusion strategy, and uses a deep cross-correlation operation to replace the simple relation operation in SiamFC, thus increasing the tracking accuracy. SiamMask [21] unified tracking and segmentation, greatly improving tracking accuracy. SiamBAN [22] optimized SiamRPN++, which improved the tracker performance by removing the previous anchor and introducing dilated convolution. SiamRN [23] proposed a novel Relation Detector(RD) structure, which can make the network filter out the interference factors in the background. Simultaneously, it also proposed a Refinement Module(RM) structure based on the coarse detection structure to achieve a more accurate tracking effect. However, the majority of the backbone networks used by these Siamese trackers are ResNet50, which can help us extract the main features effectively in most scenes. However, this study was originally intended to use ResNet101 as the backbone network to extract features because target recognition and tracking at sea are more difficult and require more characteristics. However, a more sophisticated backbone network will eventually increase the network parameters and calculation requirements, slowing the pace of the model. Finally, we introduce a model compression method for knowledge distillation after careful consideration. The newly proposed SiamKD method, when compared to the existing Siamese tracking algorithms, not only extracts rich features for target identification and positioning but also makes the model lightweight and enhances model performance.

B. BOUNDING BOX PROBABILITY DISTRIBUTION

Several articles on bounding box probability distribution modeling with a focus on target detection have recently been published. To address the issue of increasing the position accuracy of the target prediction box, He et al. [24] created a new network structure to forecasts not only the position coordinates but also the position variance for all boxes. The experimental findings demonstrate the accuracy of the algorithm. Choi et al. [25] also examined the uncertainty of the bounding box. Without altering the number of calculations or the network structure of YOLOv3, it is expected that the positioning accuracy of the box would increase true positive (TP) and decrease false positive (FP) while still operating at breakneck speed. Meyer [26], on the other hand, focused on the turning point in Huber loss. After theoretical derivation, the study shows that the turning point selected in the current Smooth L1 loss is a problem. The author of Generalized Focal Loss [27] attempted to represent a general probability distribution because it is more adaptable and can handle complex data in the actual world better. However, probability distribution modeling has rarely been applied to target tracking. Presently, bounding box prediction using existing trackers has only four output values, which is equivalent to optimizing a Dirac delta distribution. It does not account for the ambiguous area of the bounding box labels and does not estimate uncertainty. However, the probability distribution of the actual data should be arbitrary; therefore, this study uses object detection probability distribution modeling to obtain a general probability distribution, which represents the uncertainty degree of bounding box location, which is then used to increase the model's accuracy.

C. KNOWLEDGE DISTILLATION

In recent years, convolutional neural networks have achieved remarkable results in computer vision, including image classification [28], object detection [29], and semantic segmentation [30], thanks to continuous improvements in datasets and computing unit performance. However, network performance is typically inversely correlated with network structure complexity. Consequently, the more complex the network structure, the deeper and wider the model, and the better the network performance. However, as the number of parameters and the number of calculations increases in tandem, the speed of the model inevitably decreases. Deep neural networks have a large compression space, as demonstrated by the fact that approximately half of their weights do not affect the performance of the network [31]. The model's capacity for generalization was diminished by overfitting caused by excessive parameters. The compression model problem has been successfully addressed by knowledge distillation, which accelerates network training and enhances the model performance. The instructor model imparts knowledge to

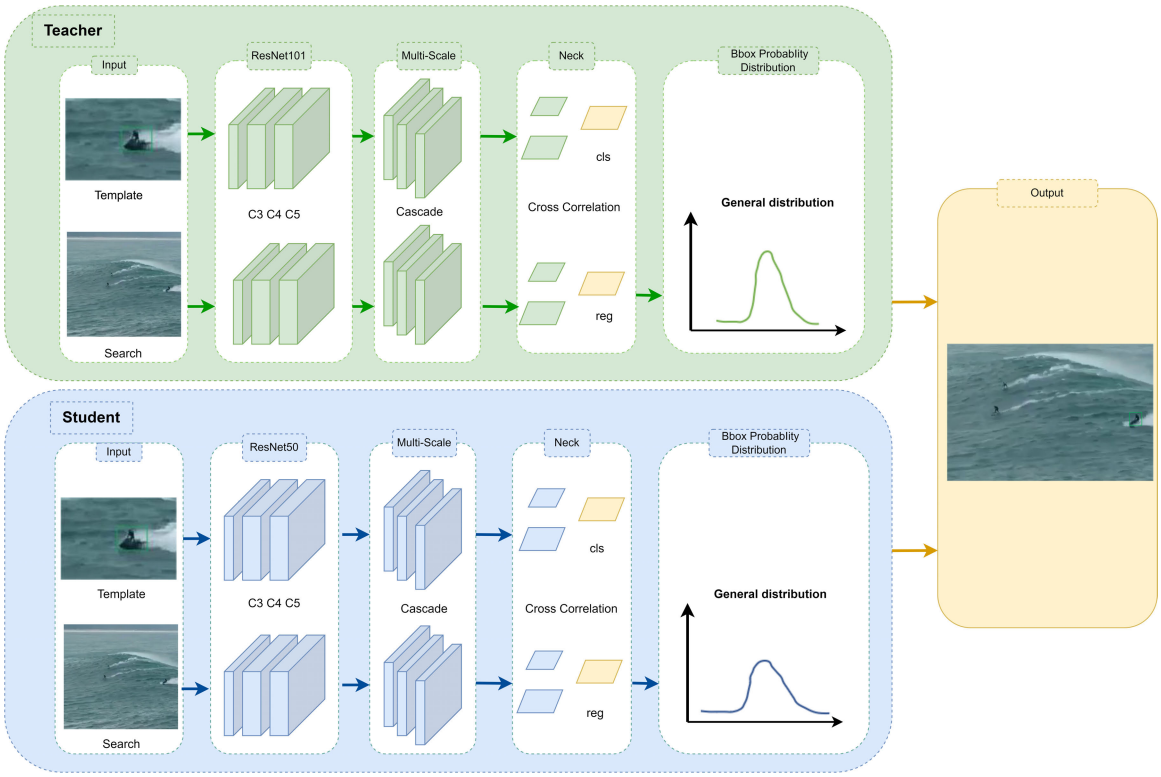


FIGURE 1. The flowchart of the SiamKD framework.

the student model. Since AI Godfather Hinton [32] developed an information distillation method, which has piqued the interest of academics, who demonstrated that network compression has significant research value [33]. Although other fields, such as semantic segmentation, object identification, and image classification, have widely used the method of information distillation, the topic of tracking has largely been left unexplored. More importantly, because this study employs ResNet101 as the foundational network to extract features, the number of parameters and calculations in the model skyrockets. Additionally, the approach in this study was simplified through knowledge distillation.

III. OUR PROPOSED SIAMKD ALGORITHM

Poor tracking results are caused by interference elements, including waves and floatable objects, according to an investigation conducted in challenging ocean environments. To reduce the fuzziness of the bounding box and improve its localization accuracy, this study introduces a bounding box probability distribution. In order to match the performance of the teacher network and extract as many features as possible, ResNet50 was utilized as the student network, while ResNet101 was used as the teacher network for pre-training the model.

A. OVERALL NETWORK FRAMEWORK OF SIAMKD

A flow diagram of SiamKD is shown in Fig. 1. This study uses SiamBAN as our benchmark, and the overall flow of the network is similar to that of SiamBAN. To pre-train the model, ResNet101 was used as the backbone network for the teacher network, and ResNet50 was used for the student network portion. The image is input into the backbone network as a template image and searched, extracting features from conv3, conv4, and conv5. Correlation convolution is then performed on the features of the template and search branches, resulting in three feature maps for classification and three feature maps for regression. Furthermore, the classification and regression feature maps are fused using an averaging-specific fusion algorithm to create the final feature maps. Finally, the student network is employed to match the performance of the teacher network, and the softmax function is used to describe the output of the regression section as a probability distribution. The student network is used in this study to match the performance of the teacher network, as the teacher network has already been trained.

B. SIAMBAN

The majority of current Siamese trackers rely on multi-scale searches or preset anchor boxes to precisely determine the scale and proportion of the target. Nevertheless, they frequently require intricate heuristic arrangements. SiamBAN suggests a straightforward but efficient tracking method to examine the expressive potential of Fully Convolutional Networks(FCN). With a unified FCN that directly categorizes foreground and background while regressing the target box,

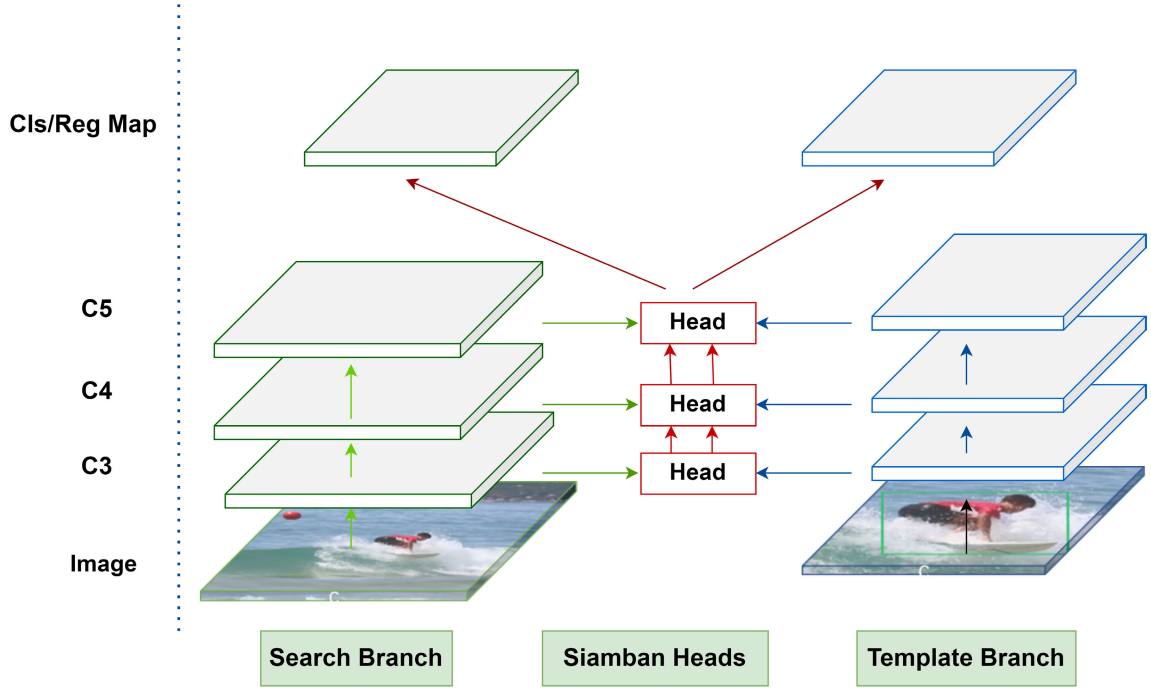


FIGURE 2. The representation of SiamBAN framework.

SaimBAN views visual tracking as parallel classification and regression problems. This box-free design, which removes the hyperparameters associated with candidate boxes, makes the SiamBAN more adaptable and versatile. The primary structure is illustrated in Fig. 2.

The SiamBAN consists of several adaptive box heads and the Siamese network backbone, as shown in Fig. 2. The Siamese network backbone calculates the convolutional feature maps of the template and search blocks using a pre-trained convolutional network. The adaptive box head consists of two modules: a classification module and a regression module. The classification module classifies each point in the relevant layer as foreground or background and predicts the bounding boxes for the corresponding position using the regression module.

C. BOUNDING BOX PROBABILITY DISTRIBUTION

Conventional bounding box prediction only has four outputs. This is equivalent to maximizing the Dirac delta distribution for each output, which is a probability distribution with an integration of 1 over a specified interval. In other words, a supervised signal is present at one site but not at the other sites. The specific mathematical premise is that the relative offsets of the coordinates to the four sides of the bounding box are utilized as regression targets, and the bounding box regression models the label y as a Dirac delta distribution, $\delta(x-y)$ satisfying $\int_{-\infty}^{+\infty}\delta(x-y)dx = 1$(x is the input value), which is typically achieved by using a fully connected layer. Formally, label y can be expressed in the following form:

$$y = \int_{-\infty}^{+\infty}\delta(x-y)dx. \tag{1}$$

It is evident from (1) that higher certainty estimation is necessary because the standard Dirac delta distribution is overly strict. It is challenging to precisely locate a target in a complex ocean environment using this distribution. On the other hand, the probability distribution of real data can be arbitrary. Therefore, unlike the previous Dirac delta distribution, this study employs the bounding box probability distribution to detect objects and directly learns the possible general distribution $P(x)$ without the use of additional priors. Given that the label range of y is: $y_0 \leq y \leq y_n$, we can obtain the predicted value y_1 from the model, where $y_0 \leq y_1 \leq y_n$.

$$y1 = \int_{-\infty}^{+\infty} P(x)xdx. \tag{2}$$

The continuous domain integral is transformed to a discrete representation through the discrete range $[y_0, y_n]$ to generate a set $y_0, y_1, \ldots, y_i, y_{i+1}, \ldots, y_n$ which is consistent with the convolutional neural network, interval $\triangle(\triangle = 1)$. As a result, it is possible to determine the discrete distribution property $\sum P(y_i = 1)$ $(0 \leqslant i \leqslant n)$ and express the estimated regression value as:

$$y1 = \sum P(y_i y_i). \tag{3}$$

The $softmaxS(.)$ layers composed of $n+1$ units can be used to obtain $P(x)$. Fig. 3 and 4 display schematic diagrams of the two distributions.

Projected bounding box labels are more appropriate for complicated situations, and the general distribution is more flexible and arbitrary than the Dirac delta distribution.

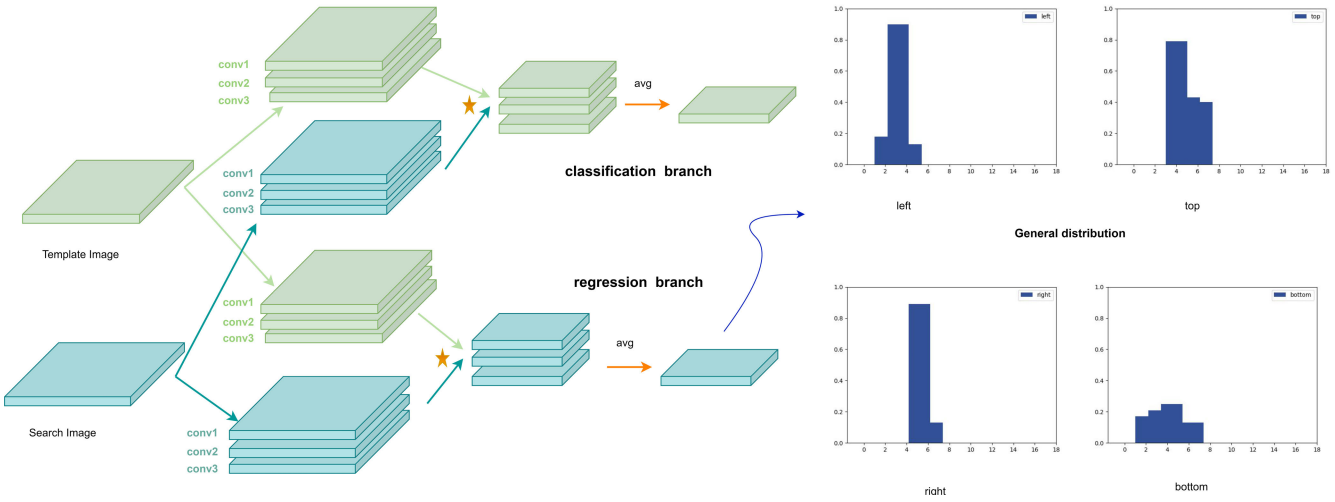


FIGURE 3. The representation of Dirac delta modeling.

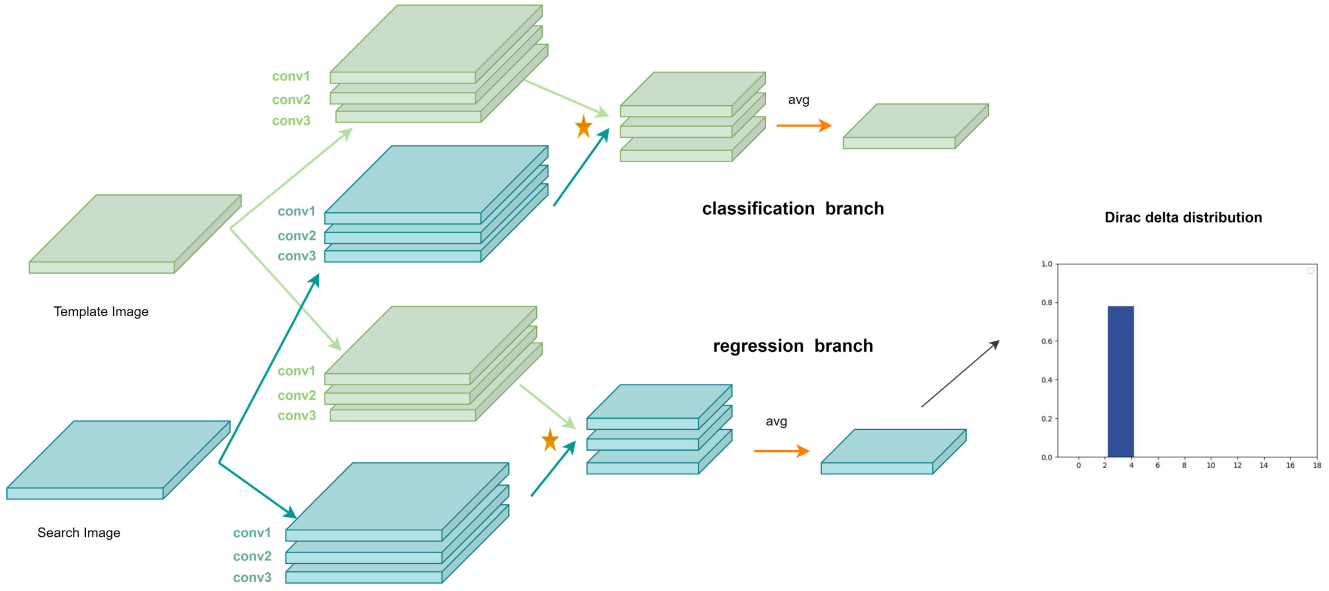


FIGURE 4. The representation of General distribution modeling.

D. KNOWLEDGE DISTILLATION

The knowledge distillation procedure consists of three stages: first, training a large model; second, calculating a ''soft target'' based on the output of the large model; and third, training a small model using both the soft target and the original hard target. During the large model stage, the network is trained in the same way as an ordinary network using regular sample labels. Once a well-performing model is obtained, the soft target is calculated based on its output. Finally, the small model is trained using both the hard target and the soft target, with the parameters adjusted accordingly.

Knowledge distillation is primarily used in the location section of this study. Probability modeling of the bounding box was used to obtain the general distribution of the bounding box. The bounding box has 4n logical values because one of its sides has n logical values. Each logical value acts on a softmax function with temperature t. The student's positioning knowledge is softened by allowing the probability distribution of the student's bounding box to fit the teacher's probability distribution. The following are the precise steps in knowledge distillation.

The training process for the large model is identical to that of the SiamBAN, except that the backbone network is replaced with ResNet101 and the regression part models the sample label as a general probability distribution. The large model is trained on samples and sample labels, which are modeled as general probability distributions.

Calculate the ''soft target'' phase. We obtained the predicted value of the bounding box tag from the previous phase,

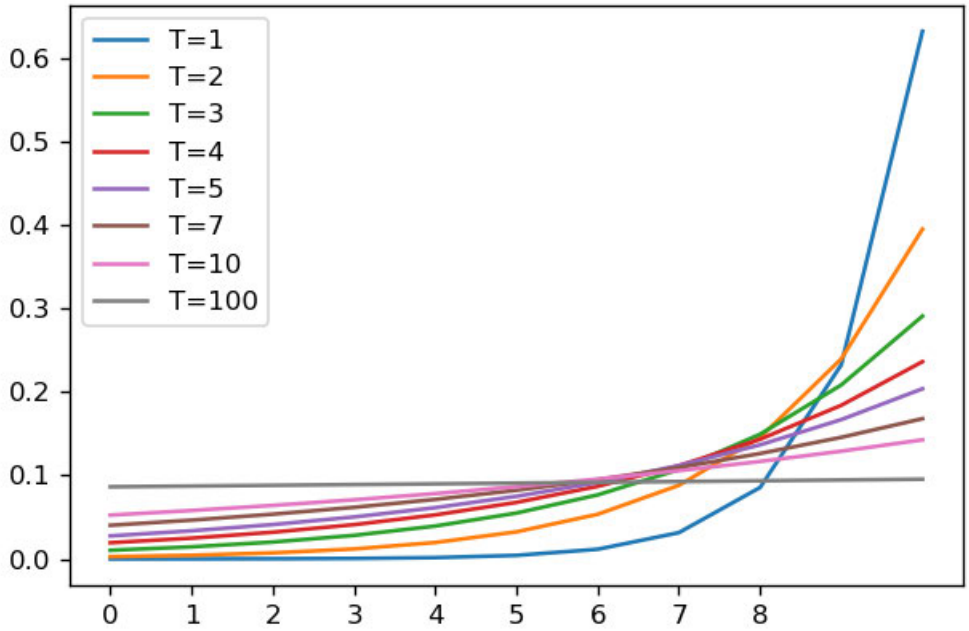


FIGURE 5. Knowledge distillation temperature visualization experiments.

and then use the (4) to calculate a "soft target".

$$q_i = \frac{exp(y_i/t)}{\sum exp(y_i/t)} (0 \leqslant y_i \leqslant n). \tag{4}$$

In (4), q_i represents soft target value and t represents distillation temperature. The larger the value of t, the more knowledge can be learned, but at the same time, the more noise there is. Therefore, when designing the network, the parameters must be adjusted to determine the appropriate temperature. The effects of the visualization experiments on different values of t, including $t = 1, 2, 3, 4, 5, 7, 10, 100$, are shown in Fig. 5. We discovered that suitable performance occurred at $t = 5$ (the balance between noise and learned knowledge); therefore, we set $t = 5$.

Train the small model. The backbone network remains the same as SiamBAN, and the regression part, similar to the large model, models the sample labels as a general probability distribution. The small model was trained by continuously adjusting the parameters of the "hard targets" and "soft targets" to fit the performance of the large model.

The knowledge distillation framework is shown in Fig. 6.

E. LOSS FUNCTION

The total loss function in the SiamBAN network is written as follows:

$$L = \lambda_1 L_{cls} + \lambda_2 L_{reg}. \tag{5}$$

In (5), L_{cls} is cross-entropy loss, L_{reg} is IoU loss, $\lambda_1 = \lambda_2 = 1$. The bounding box label is modeled as a general distribution in this study, and the bounding box probability distribution is described. To optimize its shape, the DFL loss is introduced to replace the IoU loss in the regression. This improves the shape of $P(x)$ and forces the network to quickly focus on values close to the label by increasing the probabilities of y_i and y_{i+1} (the closest values to y, where $y_i \leqslant y_{i+1}$). Thus, the regression loss function becomes:

$$\begin{aligned} L_{reg} =& DFL(S_i, S_{i+1}) \\ =& -(y_{i+1} - y)log(S_i) + (y - y_i)log(S_{i+1}). \end{aligned} \tag{6}$$

where $S_i = \frac{y_{i+1}-y}{y_{i+1}-y_i}$, $S_{i+1} = \frac{y-y_i}{y_{i+1}-y_i}$.

Furthermore, to match the performance of the student network with that of the teacher network, this study combines the general loss function formulas of knowledge distillation:

$$Totalloss = \lambda softloss + (1 - \lambda)hardloss. \tag{7}$$

where softloss is produced by the teacher and student networks, and hardloss is the conventional cross-entropy loss employed in conventional models. Here, we set $hardloss = L$. The total loss function of soft loss should be equal to the sum of the loss functions of the four edges because the bounding box has four edges. The loss functions of the four edges are as follows:

$$l_{left} = l_{right} = l_{top} = l_{bottom} = -\frac{1}{n}\sum y_i log(p_i). \tag{8}$$

where the teacher network predicts the label for the bounding box as y, and the student network predicts the label as p.

$$softloss = l_{left} + l_{right} + l_{top} + l_{bottom}. \tag{9}$$

To summarize, the student network's total loss function formula is as follows:

$$totalloss = \lambda(l_{left} + l_{right} + l_{top} + l_{bottom}) + (1 - \lambda)L. \tag{10}$$

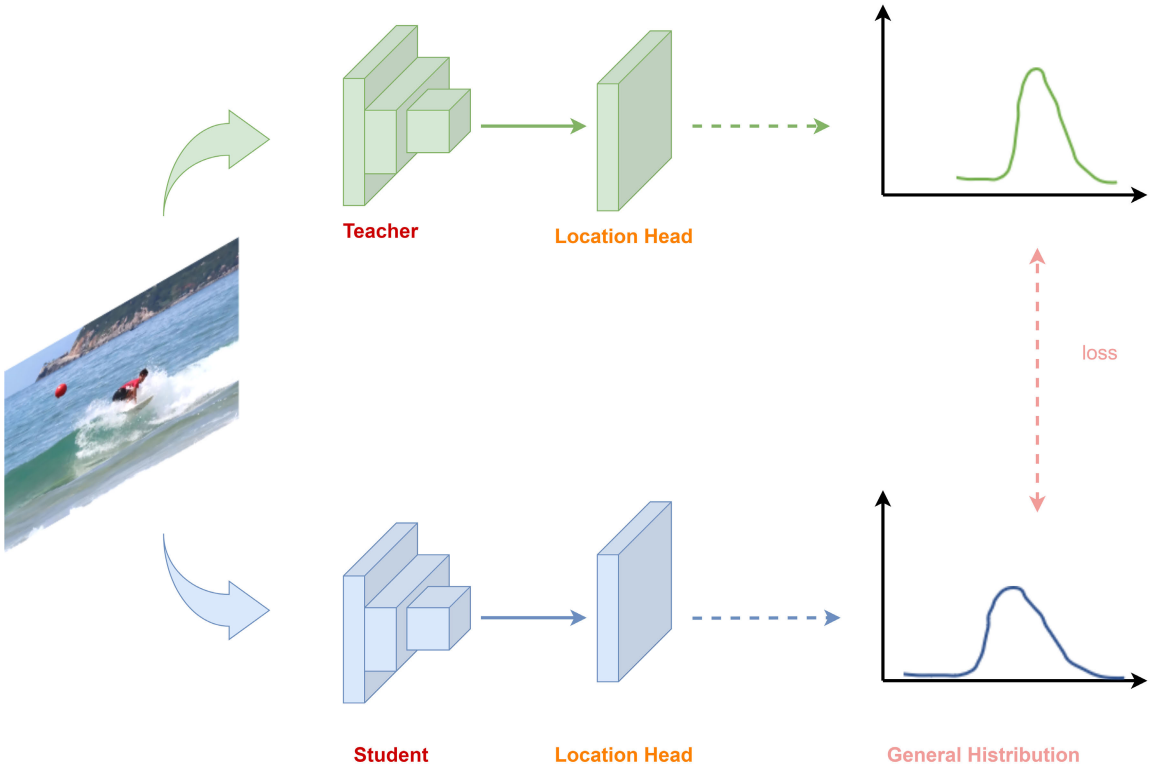


FIGURE 6. The schematic diagram of the Knowledge Distillation framework.

where $\lambda = 0.5$. The total loss function of the teacher network is L.

IV. EXPERIMENT

A. EXPERIMENTAL DETAILS

The experimental setup for this study was Pytorch 1.10.0, 1 * 24G RTX3090 GPU, Python 3.8, CUDA 11.3.

In this study, we initialized the backbone network using weights pre-trained by ImageNet and froze the parameters of the top two layers of knowledge, similar to the SiamBAN experiment. Our network was trained on stochastic gradient descent (SGD) with minibatch 28. This research trained a total of 20 epochs, with a warm-up learning rate of 0.001-0.005 for the first 5 epochs and 0.005-0.00005 for the last 15 epochs. The exponential decay was 0.00005. Only the adaptive box head was trained in the first ten periods of this study, and the backbone network was fine-tuned with a current learning rate of 0.1 in the final 10 epochs. The momentum and weight decay were set as 0.9 and 0.0001. This study used ResNet101 as the structural backbone for the teacher network and ResNet50 for the student network.

B. EXPERIMENTAL DATASETS

The training datasets used in this study were GOT10K and LaSOT. GOT10k consists of 11668 videos divided into five main categories with a total of 563 categories. LaSOT comprises 1400 video sequences organized into 70 categories. We used these two datasets to train the model to improve its generalization ability. Simultaneously, the marine dataset labeled in this research, as well as several open datasets, including VOT2018, VOT2019, OTB100, and NFS, were employed as the test set to verify the efficacy of the algorithm.

The marine dataset used in this study includes 100 video sequences that predominantly feature two types of ships and humans. The initial data for the video sequences in this dataset were obtained from the Singapore Maritime Dataset, UAV, and web crawler. The dataset was annotated using the LabelImg tool, and the annotation format was compatible with the OTB100 datasets. Some of the data are shown in Fig. 7.

C. EVALUATING INDICATOR

1) OTB100

a: PRECISION

Euclidean distance between the central point of the prediction box and the ground truth central point. (11) shows the mathematical representation.

$$p = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \tag{11}$$

where (x_1, y_1) are the coordinates of the center point of the prediction box, and (x_2, y_2) are the coordinates of the center point of the ground truth.

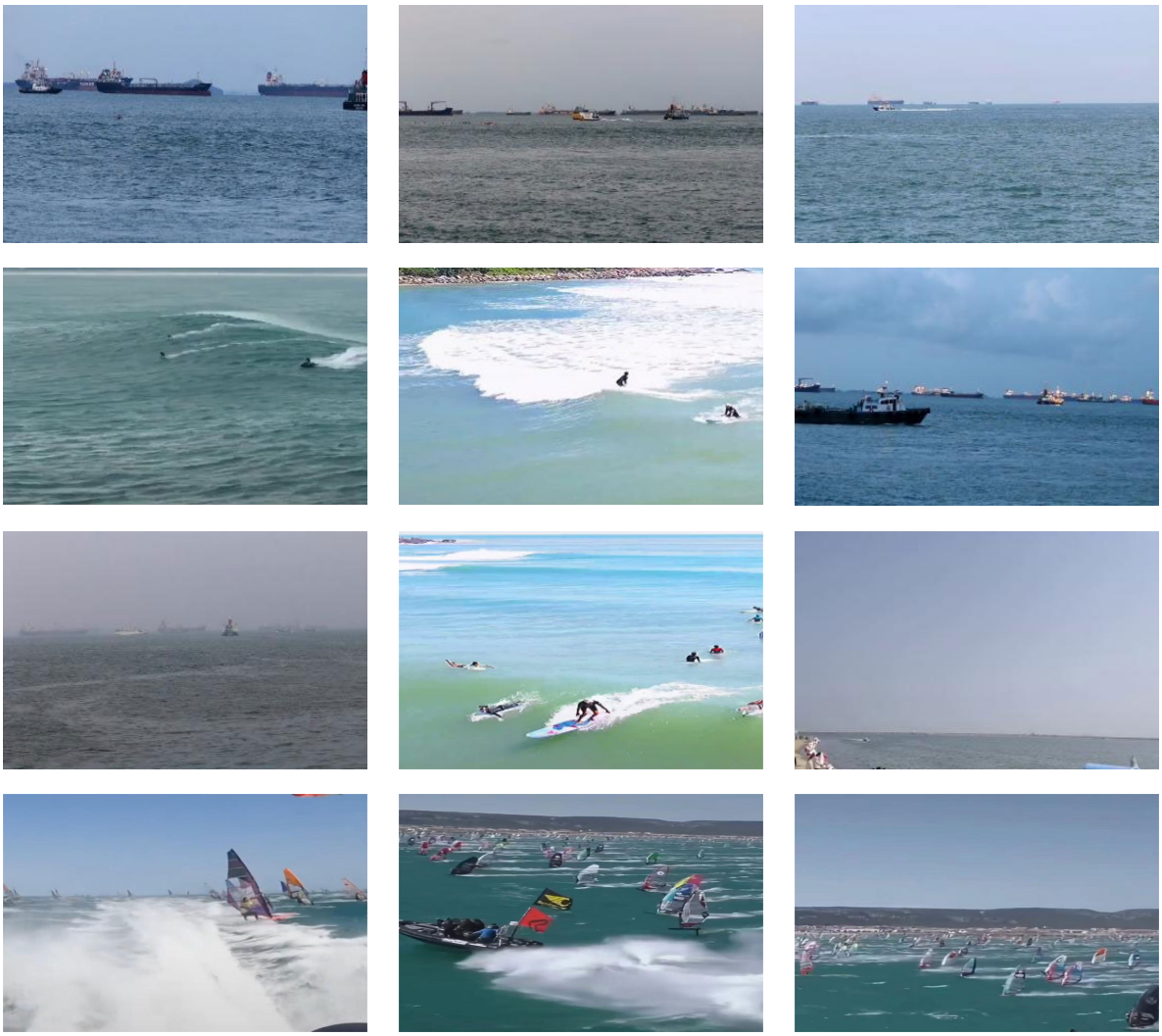


FIGURE 7. The display of Sea dataset partial images.

b: SUCCESS RATE

The Intersection over Union (IoU) of the prediction box to the pixels in the ground truth region. (12) shows the mathematical representation.

$$s = |\frac{BB_{tr} \cap BB_{gt}}{BB_{tr} \cup BB_{gt}}|. \tag{12}$$

where BB_{tr} is the predicted box region pixel and BB_{gt} is the ground truth region pixel.

2) VOT

a: ACCURACY

To evaluate the tracker's accuracy, the larger the value, the higher the accuracy. The accuracy of a sequence's t-frame is defined as (13), borrowing from the IoU definition.

$$\varphi_t = \frac{A_t^G \cap A_t^T}{A_t^G \cup A_t^T}. \tag{13}$$

where A_t^G represents the t-frame bounding box corresponds to the ground truth and A_t^T represents the tracker's predicted bounding box in a t-frame. More precisely, $\varphi_t(i,k)$ is the accuracy of the i-th tracker on the t-frame in the k-th repeat. The number of repetitions was set to N_{rep}, and the accuracy of the t-frame can be defined as follows:

$$\phi_t(i) = \frac{1}{N_{rep}} \sum \phi_t(i,k)(1 \leqslant k \leqslant N_{rep}). \tag{14}$$

The i-th tracker's average accuracy is defined as:

$$\rho_A(i) = \frac{1}{N_{valid}} \sum \phi_t(i)(1 \leqslant t \leqslant N_{valid}). \tag{15}$$

where N_{valid} is the number of valid frames.

b: ROBUSTNESS

To evaluate the stability of the tracker, the larger the value, the better the stability. Similar to how accuracy is defined, let $F(i,k)$ be the number of failures of the i-th tracker in the k-th

TABLE 1. Performance results of SiamKD and other advanced trackers on the sea dataset.

	SiamRPN	SiamRPN++	SiamMask	ATOM	DIMP	SiamFC++	PrDiMP	SiamBAN	SiamKD
Success rate	0.655	0.690	0.678	0.703	0.742	0.754	0.759	0.691	0.684
Precision	0.856	0.902	0.897	0.708	0.757	0.745	0.754	0.899	0.904

TABLE 2. Performance values of SiamKD and other advanced trackers on VOT2018.

	RCO	UPDT	SiamRPN	ATOM	SiamRPN++	SiamBAN	SiamKD
EAO	0.376	0.379	0.384	0.401	0.417	0.452	0.451
Accuracy	0.507	0.536	0.588	0.590	0.604	0.597	0.601
Robustness	0.155	0.184	0.276	0.203	0.234	0.178	0.181

TABLE 3. Performance values of SiamKD and other advanced trackers on VOT2019.

	SPM	SiamRPN++	SiamMask	DIMP	SiamBAN	SiamKD
EAO	0.275	0.285	0.287	0.321	0.327	0.329
Accuracy	0.577	0.599	0.594	0.581	0.602	0.602
Robustness	0.507	0.482	0.461	0.371	0.396	0.401

repetition. The average robustness of the i-th tracker is then defined as:

$$\rho_R(i) = \frac{1}{N_{rep}} \sum F(i, K)(1 \leqslant k \leqslant N_{rep}). \tag{16}$$

c: EAO

Reflects the overall tracker performance. It is defined as follows: if there is a frame video, the tracker's coverage accuracy on this video is the average value of each frame's accuracy, expressed in ψ. The precise formula is as follows:

$$\psi_{N_s} = \frac{1}{N_s} \sum \psi_i (1 \leqslant i \leqslant N). \tag{17}$$

An ideal EAO is to average ψ_{N_s} corresponding to N_s from N_{low} to N_{high}, that is, the expected average coverage.

3) NFS

AUC, namely accuracy, has the same definition as in the VOT dataset.

D. COMPARISON WITH OTHER TRACKERS

The performance of the SiamKD tracker proposed in this study is compared with that of several existing advanced trackers on several benchmark datasets and the dataset annotated in this study. The tracker used in this study performed excellently.

The sea dataset used in this study is composed of various videos, and the details are shown in Section B. The evaluation metrics used in our study are consistent with those of OTB100 since the annotation format is the same. The performance of SiamKD and several advanced trackers on this dataset is presented in Table 1. Our approach exceeds previous algorithms in terms of precision.

The VOT2018 benchmark consisted of 60 progressively more challenging sequences. The results of SiamKD and other advanced trackers on this dataset are presented in Table 2. Among the approaches in Table 2, SiamBAN achieved the highest EAO (0.452), while SiamRPN++ achieved the highest accuracy (0.604). Although the EAO of SiamKD is marginally lower (0.451) than that of SiamBAN, and its failure rate is higher (0.181), its accuracy is higher (0.601). Compared to SiamRPN++, SiamKD has slightly lower accuracy but a greater EAO and a lower failure rate.

VOT2019 is another dataset that is used to assess single-target trackers. The video sequences in VOT2019 were more difficult than those in VOT2018. Table 3 lists the performance of the SiamKD and other advanced trackers on this dataset. Table 3 shows that our tracker's accuracy is the same as that of the SiamBANs, which exceeds the accuracy of the other trackers listed in Table 3. Although the failure rate(0.401) was higher than that of SiamBAN and DiMP, the EAO(0.329) was higher than that of the other trackers.

The OTB100 is a popular dataset for visual object tracking, consisting of 100 annotated video sequences. The performance of SiamKD and other advanced trackers on this dataset is shown in Fig. 8. While SiamKD's performance in the full video sequence of OTB100 is not the greatest, it still outperforms the majority of the algorithms evaluated in the comparison. Additionally, our approach surpasses other algorithms in motion blur and fast motion video sequences,

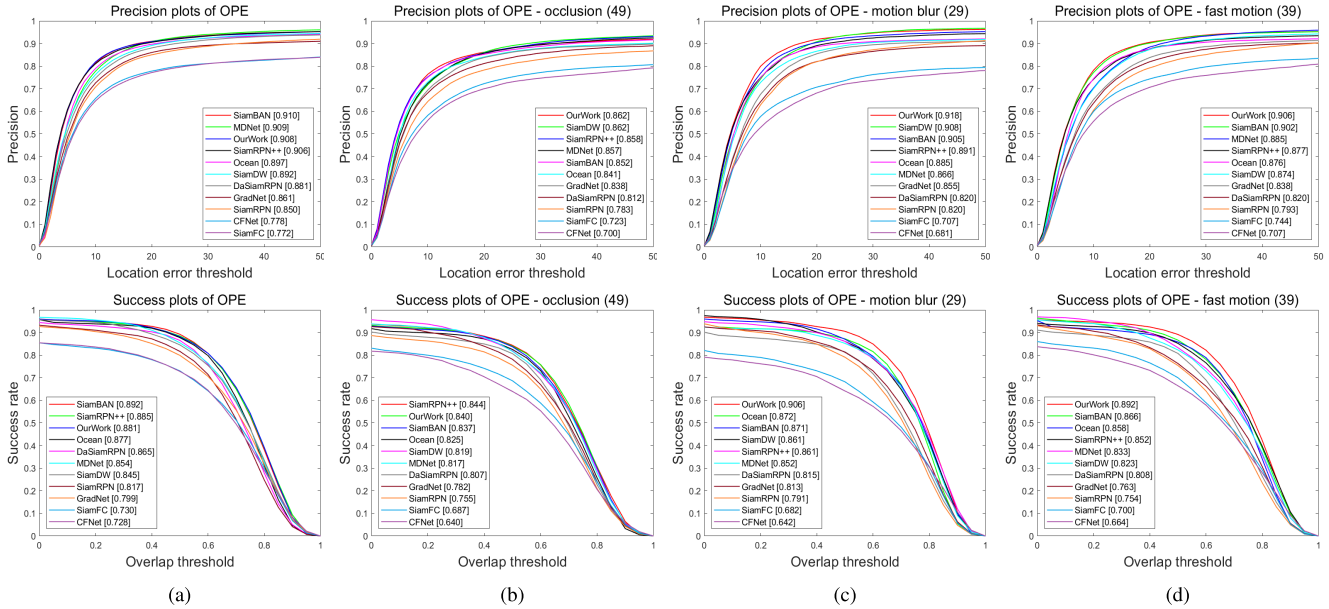


FIGURE 8. Performance results of SiamKD and other advanced trackers on OTB100, a) whole video sequences, b) occlusion video sequences, c) motion blur video sequences, d) fast motion video sequences.

TABLE 4. Performance results of SiamKD and other advanced trackers on NFS.

	MDNet	ECO	UPDT	DiMP	SiamBAN	SiamKD
AUC	0.422	0.466	0.537	0.620	0.594	0.612

TABLE 5. Parameter of SiamKD and other several models.

	SiamRPN	MDNet	DiMP	SiamBAN	SiamKD-WithoutKD	SiamKD
Params(M)	68.54	42.3	26.1	84.6	104.7	91.4
FLOPs(G)	71.1	-	-	91.8	117.7	106.3

TABLE 6. Ablation study on VOT2018. SiamBAN is baseline. BBPD: Bounding Box Probability Distribution, KD: Knowledge Distillation.

	Baseline	Baseline+BBPD	Baseline+KD	Baseline+BBPD+KD(Ours)
EAO	0.452	0.451	0.452	0.451
Accuracy	0.597	0.599	0.600	0.601
Robustness	0.178	0.178	0.180	0.181

as shown in Fig. 8. It also performs well in video sequences with improved occlusions. The success rate was barely any lower than that of SiamRPN++.

The NFS dataset consisted of 100 videos that were captured with higher frame-rate cameras in realistic settings. The AUC of SiamKD and other sophisticated trackers in this dataset are listed in Table 4. Despite the poor performance of our tracker, its AUC(0.612) is higher than that of the first SiamBAN(0.594) and is second only to that of DiMP(0.620).

Table 5 compares the model parameters of different tracking methods before and after applying knowledge distillation. The number of parameters for several tracking algorithms is listed. Introducing the knowledge distillation method significantly reduced the parameter numbers compared to the original SiamBAN. However, the number of parameters was still higher than that of the original SiamBAN network. Despite this, our approach achieved better tracking accuracy.

E. TRACKING PERFORMANCE DISPLAY

To better illustrate our tracking performance,we display the ground truth and our tracking results in some images, as shown in Fig. 9.

F. ABLATION STUDY

We conducted an ablation study on VOT2018 to investigate the impact of individual components in SiamKD.

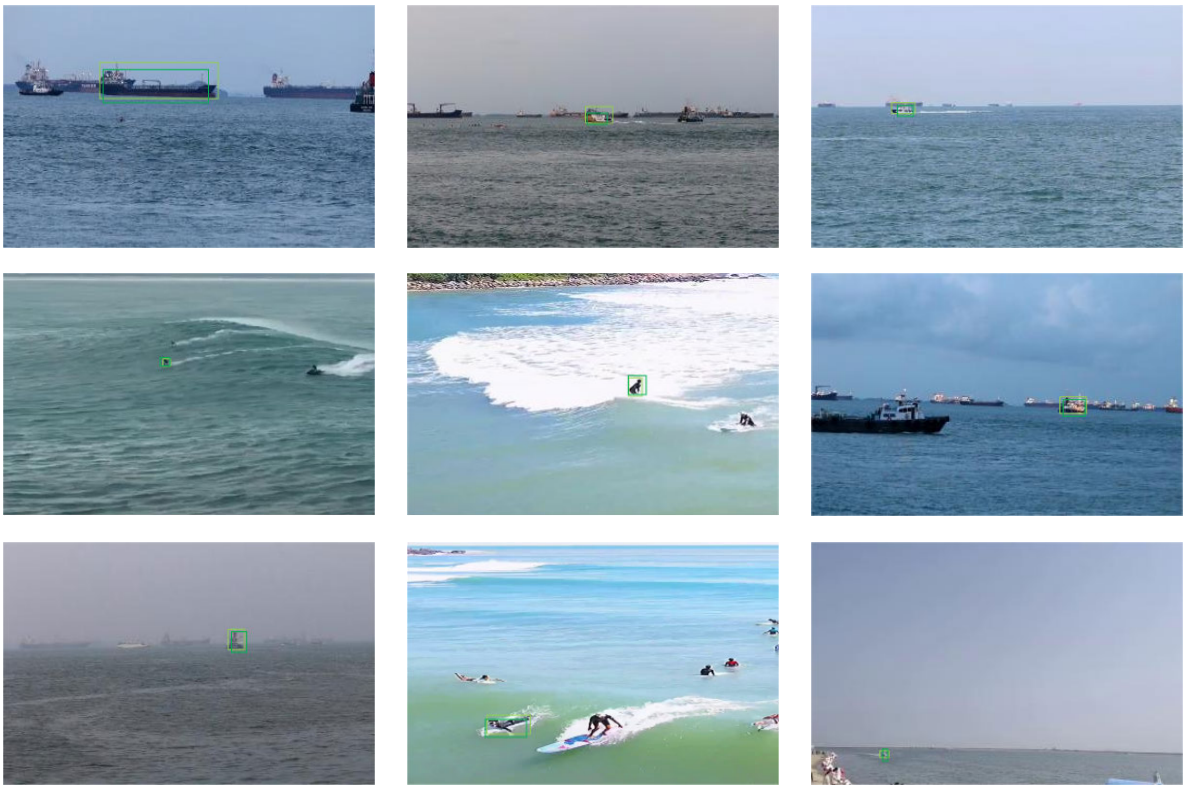


FIGURE 9. The display of tracking performance: the light green box represent ground truth; the dark green box represent our tracking results.

Using SiamBAN as the baseline, Table 6 shows that it achieved an EAO of 0.452. Although there was no significant improvement in EAO after adding the BBPD and KD modules, there was no significant decrease either. On the other hand, adding the BBPD or KD modules individually resulted in an improvement in the accuracy score from 0.597 to 0.599 or 0.600 respectively. When both modules were added to the baseline, the accuracy score improved significantly from 0.597 to 0.601. This demonstrates the importance of the proposed method in achieving higher tracking accuracy.

V. CONCLUSION

In this study, we propose a SiamKD network that models conventional bounding box regression as a general probability distribution, improving target localization accuracy and reducing the impact of distractions on tracking performance. We also introduce the knowledge distillation model compression technique, which significantly enhances the model's ability to extract richer features using more sophisticated networks while keeping the model lightweight. Our network performs well on the datasets evaluated in this study as well as on VOT2018, VOT2019, OTB100, and NFS, as demonstrated by extensive experimental results.

Despite its excellent performance, the tracking algorithm presented in this study has some limitations. Firstly, only labeled datasets are used for testing in this study, and more datasets should be labeled for model training in the future. Additionally, the removal of rain and fog should also be taken into consideration in the future as this study only addresses the presence of waves and disturbances in the ocean, not the more complex scenario of rainy and foggy weather, which is the primary cause of most maritime accidents.

REFERENCES

[1] J. Huang, "How to check the maritime affairs when the personnel on board accidentally fall into the water," *Pearl River Water Transp.*, no. 9, p. 2, Jun. 2017.

[2] W. N. Lin and L. Wang, "A comparison study on the rescue efficiency in the international waters: A case of the specific waters in the east China sea," *Mar. Sci. Bull.*, vol. 441, no. 4, pp. 438–446, Jun. 2019.

[3] W. P. Yao, "Application of UAV in life saving at sea," *China Sci. Technol. Panorama Mag.*, no. 17, p. 2, Jun. 2016.

[4] M. Munsif, H. Afridi, M. Ullah, S. D. Khan, F. A. Cheikh, and M. Sajjad, "A lightweight convolution neural network for automatic disasters recognition," in *Proc. 10th Eur. Workshop Vis. Inf. Process. (EUVIP)*, Lisbon, Portugal, Sep. 2022, pp. 1–6.

[5] Y. J. Zhang, "Research status and development trend of target tracking algorithm based on deep learning," *Modern Inf. Technol.*, vol. 8, pp. 82–85, 2021, doi: 10.19850/j.cnki.2096-4706.2021.08.024.

[6] Y. M. Zhang and J. Z. Ma, "Maritime target tracking algorithm based on convolutional features deep fusion," vol. 41, no. 1, p. 7, doi: 10.16208/j.issn.1000-7024.2020.01.042.

[7] X. Wu, Y. X. Zhong, Q. Q. Yue, and X. M. Li, "Scale adaptive sea surface target tracking algorithm based on deep learning," *J. Unmanned Undersea Syst.*, vol. 28, no. 6, pp. 618–625, doi: 10.11993/j.issn.2096-3920.2020.06.005.

[8] S. B. Shi, S. G. Wang, J. S. Zhu, and K. Zhou, "Ship target tracking method based on SiameseNet," *Ship Electron. Eng.*, vol. 40, no. 4, p. 4, doi: 10.3969/j.issn.1672-9730.2020.04.011.

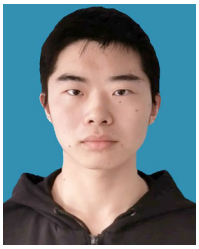
[9] Z. Zheng, R. Ye, P. Wang, D. Ren, W. Zuo, Q. Hou, and M.-M. Cheng, "Localization distillation for dense object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA Jun. 2022, pp. 9407–9416.
[10] J. Mikusinski, "On the square of the dirac delta-distribution," *Formal Power*, to be published.
[11] J. P. Gou, B. S. Yu, S. J. Maybank and D. C. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 31, pp. 1789–1819, Jul. 2021.
[12] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully convolutional Siamese fusion networks for object tracking," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Las Vegas, NV, USA, Oct. 2016, pp. 850–865.
[13] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," 2017, *arXiv:1704.06036*.
[14] Q. Wang, J. Gao, J. Xing, M. Zhang, and W. Hu, "DCFNet: Discriminant correlation filters network for visual tracking," 2017, *arXiv:1704.04057*.
[15] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic Siamese network for visual object tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Honolulu, HI, USA, Oct. 2017, pp. 1763–1771.
[16] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lack City, UT, USA, Dec. 2018, pp. 8971–8980.
[17] Z. Zhu, "Distractor-aware Siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Aug, Germany, 2018, pp. 101–117.
[18] A. F. He, C. Luo, X. M. Tian, and W. J. Zeng, "A twofold Siamese network for real-time object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lack City, UT, USA, Dec. 2018, pp. 4834–4843.
[19] Z. Zhang and H. Peng, "Deeper and wider Siamese networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4591–4600.
[20] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4282–4291.
[21] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1328–1338.
[22] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 6668–6677.
[23] S. Cheng, B. Zhong, G. Li, X. Liu, Z. Tang, X. Li, and J. Wang, "Learning to filter: Siamese relation network for robust tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Kuala Lumpur, Malaysia, Jun. 2021, pp. 4421–4431.
[24] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," 2018, *arXiv:1809.08545*.
[25] J. Choi, D. Chun, H. Kim, and H. Lee, "Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Long Beach, CA, USA, Oct. 2019, pp. 502–511.
[26] G. P. Meyer, "An alternative probabilistic interpretation of the Huber loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Kuala Lumpur, Malaysia, Jun. 2021, pp. 5261–5269.
[27] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," 2020, *arXiv:2006.04388*.
[28] Z. T. Jiang, J. Q. Qin, and S. Q. Zhang, "Parameterized pooling convolution neural network for image classification," *Pearl River Water Transp.*, vol. 48, no. 9, pp. 1729–1734, Jun. 2020.
[29] Y. Liu and Y. W. Zhan, "Survey of small object detection algorithms based on deep learning," *Comput. Eng. Appl.*, vol. 57, no. 2, pp. 37–48, Jan. 2021.
[30] Q. C. Tian and Y. Meng, "Image semantic segmentation based on convolutional neural network," *J. Chin. Comput. Syst.*, vol. 41, no. 6, pp. 1302–1313, Jun. 2020.
[31] Y. Han, "A review of knowledge distillation in deep neural networks," *Comput. Sci. Appl.*, vol. 10, no. 09, pp. 1625–1630, 2020.
[32] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[33] J. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*. New York, NY, USA: Cornell University, Nov. 2019, pp. 4793–4801, doi: 10.1109/ICCV.2019.00489.

YIHONG ZHANG (Senior Member, IEEE) received the B.S. degree in auto-control from China Textile University, Shanghai, China, in 1999, the M.S. degree in software engineering from Donghua University, Shanghai, in 2006, and the Ph.D. degree in intelligent manufacturing from The Hong Kong Polytechnic University, Hong Kong, in 2012.

In 1999, he joined the School of Information Science and Technology, Donghua University, as a Teacher. He later became an Associate Researcher and an Associate Professor, in 2013 and 2020, respectively. He has been the School's Vice President, Since 2017. His research interests include image processing, classification, target tracking, and recognition. He is a Permanent Member of the Chinese Institute of Artificial Intelligence. From 2018 to 2021, he was an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, *Neural Computing*, and *Sensors*.

QIN LIN received the B.S. degree in mathematics from the Hubei University of Science and Technology, Hubei, China. She is currently pursuing the M.S. degree in control science and engineering with Donghua University, Shanghai, China. Her research interests include single-target tracking and deep learning.

HUIZHI TANG received the B.S. degree in communication engineering from Huaibei Normal University, Huaibei, Anhui. She is currently pursuing the M.S. and Ph.D. degree in information and communication engineering with Donghua University, Shanghai, China. Her research interests include the security authentication based on privacy protection of VANET and physical layer security based on wireless communication in VANET.



YINJIAN LI received the B.S. degree in mechanical design, manufacturing and automation from the Nanjing Institute of Technology, Nanjing, China. He is currently pursuing the M.S. degree in control engineering with Donghua University, Shanghai, China. His research interests include single-target tracking and deep learning.