IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

TransCODNet: Underwater Transparently Camouflaged Object Detection via RGB and Event Frames Collaboration

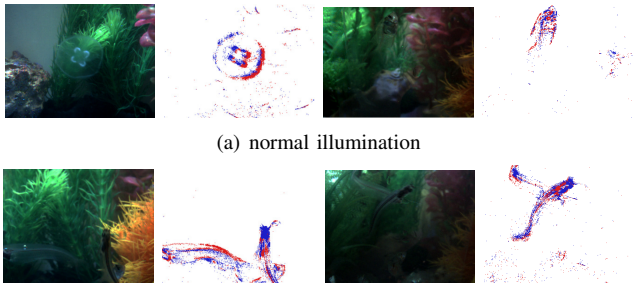
Cai Luo[1], *Senior Member, IEEE*, Jihua Wu[1], Shixin Sun[1], and Peng Ren[1], *Senior Member, IEEE*

***Abstract*—Underwater transparently camouflaged organisms can be perfectly "invisible" in the ocean to avoid the capture of predators. Due to the blurry contour boundaries of their bodies, obtaining their boundary features and determining their specific positions are challenging for detection tasks. To address this issue, first, we propose a large-scale underwater transparently camouflaged object dataset, termed Aqua-Eye, which is obtained from event data and contains five types of underwater transparent organisms, with a total of 6,497 annotated images. Second, to evaluate the effectiveness of this dataset, we propose a simple and effective detection network termed underwater Transparently Camouflaged Object Detection Network (TransCODNet), which can obtain local features and specific locations of targets, providing a better detection method for underwater transparently camouflaged organisms. In this letter, we performed ablation study and nine representative deep learning algorithms were evaluated based on the dataset. Finally, experiments show that the detection accuracy of this algorithm is 84.7%, which is superior to mainstream object detection algorithms, proving the effectiveness of the proposed method. The dataset is available at https://github.com/lunaWU628/Aqua-Eye-Dataset.**

***Index Terms*—Deep learning methods, data sets for robotic vision, object detection, event camera.**

I. Introduction

As the balancer of the marine ecological environment, the number of marine organisms accounts for 87% of the total biomass of the earth [1], among which underwater transparent organisms such as jellyfish account for a large proportion in the ocean. This type of organism is widely present in the ocean and can even cause serious impacts on marine ecosystems. Therefore, it is necessary to protect the diversity of marine organisms and monitor and recognize them. In various marine monitoring tools, machines such as Remotely Operated Vehicle (ROV) and Autonomous Underwater Vehicle (AUV) play an important role. They are equipped with visual sensors to enter the deep sea and high-risk areas to record data on marine organisms, providing the scientific basis for sustainable ocean utilization and protection.



(a) normal illumination

(b) high illumination (c) low illumination

Fig. 1. Partial images of the dataset. RGB and event frames (a) under normal illumination; (b) under high illumination; (c) under low illumination.

Living organisms on land often require pigments such as melanin to protect themselves from ultraviolet rays, while in the ocean, the difference in refractive index between water and biological tissues is relatively small, making it relatively easy to achieve body transparency [2]. Transparency is an effective method for marine organisms to achieve camouflage, and it is also a result of the natural evolution of underwater organisms. Many organisms in the ocean use methods that reduce transparency to avoid being caught by predators. At present, ROV and AUV equipped with frame based ordinary cameras have been widely used in monitoring underwater biological activities. However, using frame-based cameras to capture images of underwater organisms, especially transparent underwater organisms, still faces problems such as high-speed motion blur and low contrast, making it difficult to obtain the morphology of underwater transparent organisms. Event camera [3] is a novel type of visual sensor, and event-based visual systems have made great achievements in object detection [4]. Therefore, this article proposes using the event camera to obtain information such as the morphology of underwater transparently camouflaged organisms, so as to achieve the purpose of better observation and protection of underwater transparent organisms.

With the rapid development of deep learning, underwater object detection technology has made significant progress, and its algorithms have also received widespread attention and research, such as [7], [8], and [9]. At the same time, excellent underwater object datasets have also been proposed in papers such as [15], [16], and [17]. The remarkable progress in this field has greatly contributed to the advancement of

Manuscript received:August 18 2023; Revised October 17 2023; Accepted December 12 2023.

This paper was recommended for publication by Editor Markus Vincze upon evaluation of the reviewers' comments. This work was supported in part by the National Key R&D Program of China under Grant 2021YFE0111600 and in part by the Fundamental Research Funds for the Central Universities under Grant 22CX01004A. *(Corresponding author: Jihua Wu.)*

[1]Cai Luo, Jihua Wu, Shixin Sun, and Peng Ren are with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, Shandong 266580, China luo_cai@upc.edu.cn; wjh628523@163.com; sunshixin_upc@163.com; pengren@upc.edu.cn

Digital Object Identifier (DOI): see top of this page.

underwater organism detection, but the types and sample sizes of underwater object datasets they use are relatively small. Meanwhile, there is almost no attention paid to the research direction of underwater transparently camouflaged organisms. Therefore, based on underwater application scenarios and targets, this paper constructs an underwater transparently camouflaged object dataset and detector. The summary of our contributions is as follows:

- We present an underwater transparently camouflaged organisms dataset for the first time with event camera: Aqua-Eye, which includes five kinds of underwater transparent organisms with a total of 13578 images, and covers scenes with different light intensities and complex backgrounds.
- We utilize the weighted average strategy and VGGNet to fuse event frames with RGB at pixel level to further enhance the edge features of underwater transparently camouflaged organisms.
- We propose a large-scale object detection method termed TransCODNet, which is the first work to be performed using an end-to-end strategy for underwater transparently camouflaged object detection, to the best of our knowledge. The experiments have demonstrated that the methods proposed in this letter can achieve state-of-the-art performance in underwater transparently camouflaged object detection.

II. RELATED WORK

This section first reviews the detection strategies based on event cameras and then introduces the current methods for detecting camouflaged biology and transparent objects.

A. Event-based Object Detection

The event camera is a dynamic vision sensor inspired by biology. Its imaging principle differs from traditional cameras, which allows it to overcome the limitations of traditional cameras and have unique advantages, such as high dynamic range, high temporal resolution, and low power consumption, and to exert its powerful advantages in scenes with fast object motion or poor lighting [10].

At present, most event-based camera detection methods rely on converting event streams into event frames and using deep learning algorithms for detection. For example, Jiang *et al*. [21] input the APS and DVS channels generated by the event camera into the convolutional neural network for pedestrian detection through models such as YOLOv3, which can improve the detection speed and accuracy. Cao *et al*. [22] proposed a FAGC structure based on an attention mechanism and hard fusion strategy and fused event flow and grayscale frames to improve the detection accuracy of the network. Abhishek *et al*. [23] proposed a dual parallel network based on ResNet to fuse the information from frames and events, which can improve vehicle detection capabilities in driving scenarios with different weather.

Some work also utilizes the sparsity of events for detection, using Spiking Neural Networks (SNN) to improve the efficiency of processing event streams. For example, Cordone *et al*. [20] combined SNN with SSD and proposed the first SNN to be detected on a complex automotive event dataset. This method does not require a large number of timesteps, thereby improving the performance of model detection. Kugele *et al*. [25] proposed a hybrid architecture that combines the backbone of an efficient SNN based on event feature extraction with the head of an Analog Neural Network (ANN) to obtain a high-precision network, improving computational efficiency.

B. Camouflaged Object Detection

Camouflaged object detection is a new type of visual detection task, and most algorithms distinguish and locate camouflaged targets by simulating the human visual system. Camouflaged object detection based on deep learning has received widespread attention in recent years.

Fan *et al*. [11] designed the first camouflaged object detection framework called SINet, which uses deep learning methods and mainly consists of two modules: the recognition module and the search module. The recognition module uses the partial decoder component to recognize disguised creatures and the search module uses the receptive field to further search targets from complex environments. Mei *et al*. [12] proposed a bionic framework for simulating animal predatory behavior, called PFNet, which contains a global localization module and a refinement focus module to further improve the accuracy of the model. Ren *et al*. [13] proposed a camouflaged object detection approach that incorporates texture features and depth information, called TANet. This network first uses feature extractors to generate feature maps with multiple resolutions, then refines them through residual refinement block, and further enhances them through texture perception refinement block. Wang *et al*. [14] proposed a framework called D2CNet, inspired by human visual mechanisms, which includes a feature extraction module and a cross fusion module. This model simulates the process of human visual detection and achieves good results in camouflaged object detection.

C. Transparent Object Detection

Transparent objects such as glasses, glass doors, and windows are widely present in daily life, and detecting transparent objects has become popular in computer vision research recently. However, the detection of transparent creatures in nature, such as jellyfish and glass wing butterflies, has not attracted widespread attention.

At present, the most advanced method for detecting transparent objects is to use deep learning models. Specifically, Cao et al. [34] proposed the FakeMix network and improved the ASPP structure, which can dynamically obtain multi-scale features and solve the problem of boundary imbalance in transparent objects. Mei et al. [35] proposed a large-scale glass dataset (GDD) and designed a glass detection network called GDNet, which greatly promoted research on transparent objects. Jiang et al. [36] constructed a dataset called TRANS-AFF and introduced an A4T method. This method utilizes the AffordanceNet to detect transparent objects and obtain their features, and then uses a multi-step reconstruction method to gradually reconstruct the depth map of the transparent objects.

Fig. 2. (Left) Event camera. (Right) ROV equipped with an event camera.

III. THE PROPOSED DATASET

To fill in the gaps in the underwater transparently camouflaged object dataset, we constructed a large dataset called the Aqua Eye Dataset, which data is collected by the event camera. The device is shown in Fig.2. To the best of our knowledge, this is the first underwater transparently camouflaged biological dataset, which includes five representative and common organisms: flame jellyfish, moon jellyfish, X-ray fish, glass catfish, and siberian prawn.

We first use the method of [18] to convert the event camera output file (.aedat4) to RGB and event frames (.bmp) to obtain the image with a resolution of 346×260 per image. The final dataset consists of 13,578 RGB images and 13,578 event frames, of which 6,497 images were manually annotated and carefully verified by a professional labeling company. Since the different frequencies of each species appear in the video, the number of image categories also varies, with moon jellyfish having the fewest number of bounding boxes. Table I presents the quantity of bounding boxes per species.

Through statistics, it is found that the number of organisms in each picture is up to 6, and there is occlusion between organisms. Meanwhile, from the size of underwater biological images, nearly half of the targets are smaller than 10% of the image size, and these organisms can be considered small targets, with x-ray fish and siberian prawn accounting for the majority. In addition, to make the dataset more challenging and provide researchers with more diverse samples, we collected data under different illumination conditions while increasing the complexity of the background environment. Some images are shown in Fig.1. From the dataset, we found that in the event frames, the tentacles of certain organisms or small particles in water are very clear. Therefore, we hope that our dataset can also drive researchers to conduct more in-depth exploration in these fields.

Although this dataset includes complex environmental backgrounds, it cannot cover all possible scenarios, so its universality still needs to be improved. In the future, we will improve the generalization of the dataset while increasing the number of transparent organisms, so as to promote in-depth research in the field of underwater transparently camouflaged organisms.

IV. METHODOLOGY

There are two main difficulties in object detection of transparently camouflaged organisms compared to other tasks.

TABLE I
THE QUANTITY OF BOUNDING BOXES FOR EACH SPECIES

Species	Glass Catfish	X-ray Fish	Siberian Prawn	Moon Jellyfish	Flame Jellyfish
Number	4743	4744	5369	410	2232

First of all, the contrast between the target and the background is not obvious, so it is a challenge to accurately extract the edge features of the target, resulting in difficulties for the model in the process of feature extraction. The second challenge is the complex background of the underwater environment, with various interferences such as lighting and visual blur, which greatly affect detection accuracy. For these two challenges, this paper proposes two corresponding countermeasures.

A. The Fusion Network of RGB and Event Frames

In response to the issue of unclear contrast between transparent organisms and their backgrounds, this paper proposed pixel-level fusion of RGB and event frames in the dataset to supplement the edge information of underwater transparently camouflaged objects. The fusion method used was proposed by Li *et al*. [24], and the detailed diagram is shown in Fig.3. Using a deep learning framework: VGGNet, they came up with a simple and effective fusion method. This method decomposes the source image into basic parts I_k^b and detailed part I_k^d. The basic part is fused by the weighted average strategy to obtain F_b, and the detailed part is multi-layer fusion using the pre-trained VGG-19 network to obtain F_d. Finally, the basic part and the detailed part are reconstructed to obtain a fused image. Fig.3 clearly shows that the edge characteristics of the merged underwater organisms have been enhanced.

B. The Network of Underwater Transparently Camouflaged Object Detection

In response to the complex underwater environment and low detection accuracy, this paper constructs an underwater transparently camouflaged object detection framework called TransCODNet. This network combines MobileNet-V3, VarifocalLoss, and proposes a Visual Enhancement Module (VEM) to simulate human visual mechanisms. The VEM significantly improves the capability of the network to search and extract underwater transparently camouflaged objects features. Fig.4 shows the overall structure of the model.

MobileNet-V3. The feature extraction model used in the backbone part of YOLOX[5] is CSPDarkNet, and the main structure is CSPNet[6], which is an excellent network. However, the complexity of this structure makes it unsuitable for scenarios where computing resources and memory are limited on underwater platforms.

Therefore, this paper replaces CSPNet with MobileNet-V3. MobileNet-V3 adopts deepwise separable convolutions, inverted residuals, and linear bottleneck structures to construct lightweight feature extractors. In addition, it adds the attention model called SENet to allow the network to automatically learn the weights of each channel and achieve higher performance with less computational overhead. And MobileNet-V3 proposes an h-swish activation function to reduce the amount

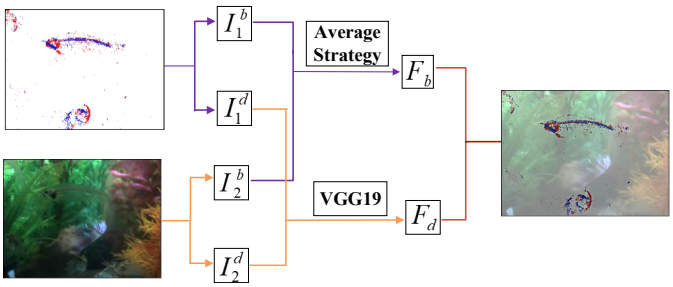


Fig. 3. The fusion process of RGB and event frames.

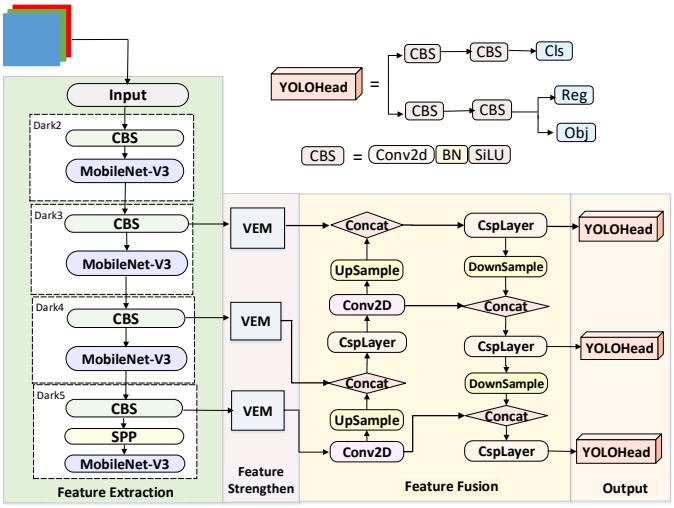


Fig. 4. The overall framework of TransCODNet.

of computation. The basic network units of MobileNet-V3 are shown in Fig.5.

The input features are first upsampled by 1×1 convolution, and the computation of the network is reduced by using 3×3 deep separable convolution. Subsequently, SENet is added to allow the network to learn feature weights based on loss, making the effective feature map weighty and focusing on the target features. The features after global average pooling are successively activated by the fully connected layer of ReLU and $hard-\alpha$ in the SE module. And finally, the 1×1 convolution is applied to decrease the dimensionality of the features. MobileNet-V3 is available in two versions, small and large, and this paper selects the small structure for experimentation. Experiments have shown that using MobileNet-V3 instead of CSPNet can improve accuracy while reducing the extra computing cost.

Visual enhancement module. Considering the similar texture features between camouflaged objects and the surrounding environment [11], using a general feature extraction structure may not accurately extract the required features. Therefore, after using the MobileDarkNet to extract features, to better extract features and obtain a larger perception field, we propose the VEM to enlarge the target area and further recognize the contour features of the target. This pattern resembles the working mode of human eyes and is a simple and useful detection method.

According to the research of Liu *et al*. [26], the Receptive Field Block (RFB) is composed by integrating multiple branches with convolution kernels and expansion convolutions of different scales. And RFB strengthens the deep features learned by the network from the lightweight CNN model and makes the detection faster and more accurate. Therefore, VEM was constructed inspired by RFB. As shown in Fig.6, the module contains 5 branches, and the size of the first convolutional layer (CBS) in each branch is 1×1, which is used for dimensionality reduction operations. Among them, the convolutional layer of branch 1 has a max-pooling layer in front of it to reduce the network complexity. After 1×1 CBS operation, branch2-4 use convolution layers of $(2i-1)\times1$ and $1\times(2i-1)$ to replace $(2i-1)\times(2i-1)(i\in2,3,4)$, which can reduce model parameters and ensure that model performance does not decrease, and then add dilated convolution to increase the receptive field, and the dilation rate is set to (2i -1). The model can simulate the role of group receptive field regions in the human visual system, and further strengthen and extract the input features.

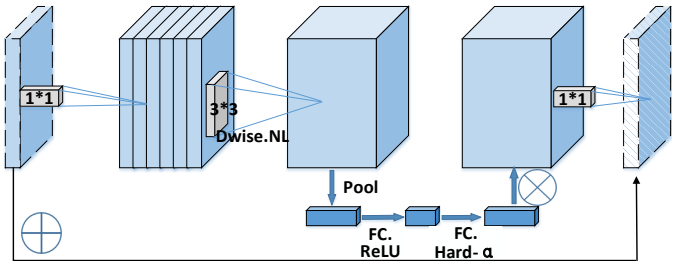


Fig. 5. Basic network unit of MobileNet-V3.

Loss function. The loss function of one-stage generally consists of three parts: positioning loss L_{reg}, classification lossL_{cls} and confidence loss L_{obj}. The confidence loss in YOLOX uses BCELoss, which is commonly used in object detection, but its detection effect on dense objects is average. Therefore, this paper replaces BCELoss with VarifocalLoss [19]. The core idea of VarifocalLoss is to solve the problem of difficulty sample imbalance by amplifying the weight of difficult samples and reducing the weight of easy-to-classify samples. The formula is as follows:

$$VFL(p,q)=\begin{cases}-q(q\log(p)+(1-q)\log(1-p)) & q>0,\\ -\alpha p^{\gamma}\log(1-p) & q=0,\end{cases} \tag{1}$$

where α=0.75 indicates a negative sample balance factor, γ=2.0 indicates a modulation factor to reduce the contribution of simple sample loss, p is the classification score for predicting IoU perception, and q is the IoU score of the target. For positive samples, set q to IoU between the bounding box and ground-truth box. And for negative samples, the training target q is 0 for all classes.

VarifocalLoss uses the training objective q to weigh the positive samples. If the ground-truth IoU of the positive samples is large, its contribution to the loss is also significant. Therefore, the focus of training is on high-quality positive samples to obtain higher accuracy values.

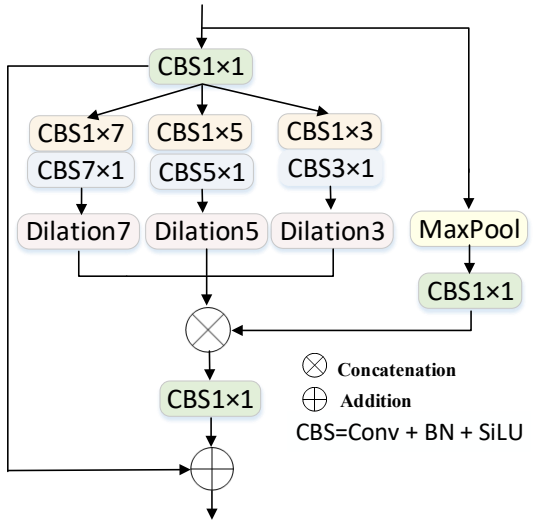


Fig. 6. Visual enhanced module.

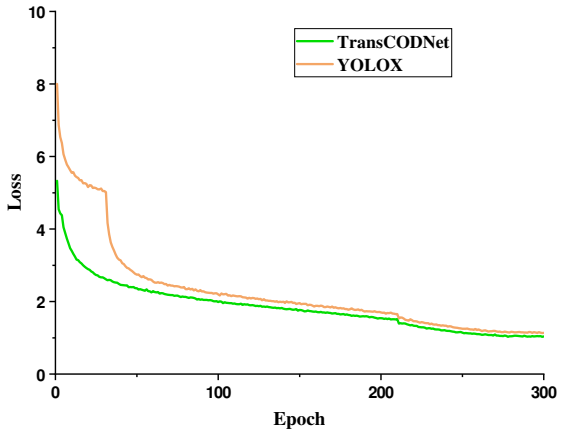


Fig. 7. Training loss curves of YOLOX and TranCODNet.

V. EXPERIMENTS AND RESULTS

A. Experiment Details

Datasets. In this paper, the Aqua-Eye mentioned above is used, and the event frames are fused with RGB according to the method of Fig.3. A total of 6497 fused images are obtained and divided into training, verification, and test sets using the ratio of 8:1:1.

Platform. The experiment runs on the Ubuntu 18.04.5 operating system and uses PyTorch as the deep learning framework on a GeForce RTX 3090 GPU. The experiment trains 300 epochs with an initial learning rate of 0.01. We adopt SGD as the optimizer and the batch size is set to 8.

Evaluation metrics. The Average Precision (AP), Mean Average Precision (mAP), the parameters of the model (Params), floating point operations per second (FLOPs), and Frames Per Second (FPS) were used as evaluation metrics to evaluate the performance of the model from the aspects of detection accuracy and efficiency. The formula for AP and mAP are as follows:

$$AP = \int_0^1 P(R)dR, \tag{2}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP(i), \tag{3}$$

where P is precision and R is recall. In general, the higher the AP and mAP, the better the detector performance.

B. Training Results

Fig.7 is the curve of the loss function. The green curve refers to the decline of the loss function for TransCODNet during training, while the yellow curve refers to the decline of the loss function for YOLOX during training.

It can be seen from Fig.7 that as the epoch increases, the loss is decreasing, and when the epoch reaches 260, the loss of the two gradually converges. Moreover, the TransCODNet has a lower loss value, with a final loss value of 1.04, while YOLOX has a final loss value of 1.13. Usually, a lower training loss results in a more accurate model, so it can be proven that the network introduced in this letter is easier to train and can converge to a more accurate model quickly.

In this experiment, TransCODNet was compared with CenterNet[27], EfficientDet[28], Faster R-CNN[29], RetinaNet[30], SSD[31], Yolov3[32], Yolov5, Yolov7[33] and YOLOX algorithms. To make the experiment more convincing, we used a 10-fold cross-validation experimental method, and the experimental results are presented in Table II. In the results of detecting fused images, except for Faster R-CNN, SSD, and Yolov3, the detection accuracy of other algorithms is higher than 80.0%. And the TransCODNet has the highest detection accuracy, reaching 84.7% mAP.

In the one-stage algorithm, CenterNet has an accuracy of 81.1%. The detection accuracy of EfficientDet is 79.8%, which is at the middle level among many detection algorithms. The accuracy of SSD is only 78.0%, and RetinaNet is 80.0%. Due to the high speed that CentetNet, EfficientDet, and RetinaNet can achieve during the experimental process, they can be used in situations where computing resources are limited. In the YOLO series of algorithms, except for Yolov3, which has the mAP of only 71.4%, the accuracy of Yolov5, Yolov7, and YOLOX are not significantly different, with 83.2%, 83.7%, and 83.6%, respectively. Faster R-CNN is representative of the two-stage algorithm, with an accuracy of 77.7%. In practical experiments, this algorithm has a slow detection speed and low accuracy, so it is not recommended for the detection of underwater transparently camouflaged organisms.

We present the results of using TransCODNet for detection in Fig.8 to better show our dataset and experimental results. From the graph, it can be seen that the accuracy of using the algorithm proposed in this article to detect fused images is relatively high, and the accuracy value obtained by detecting event frames is the lowest. However, in poor lighting conditions, the detection accuracy of event frames is higher.

C. Ablation Study

To analyze the effect of the suggested improvements on model performance and the difference in detection accuracy between fused images and RGB images, we conducted ablation experiments. In the ablation experiment, we trained

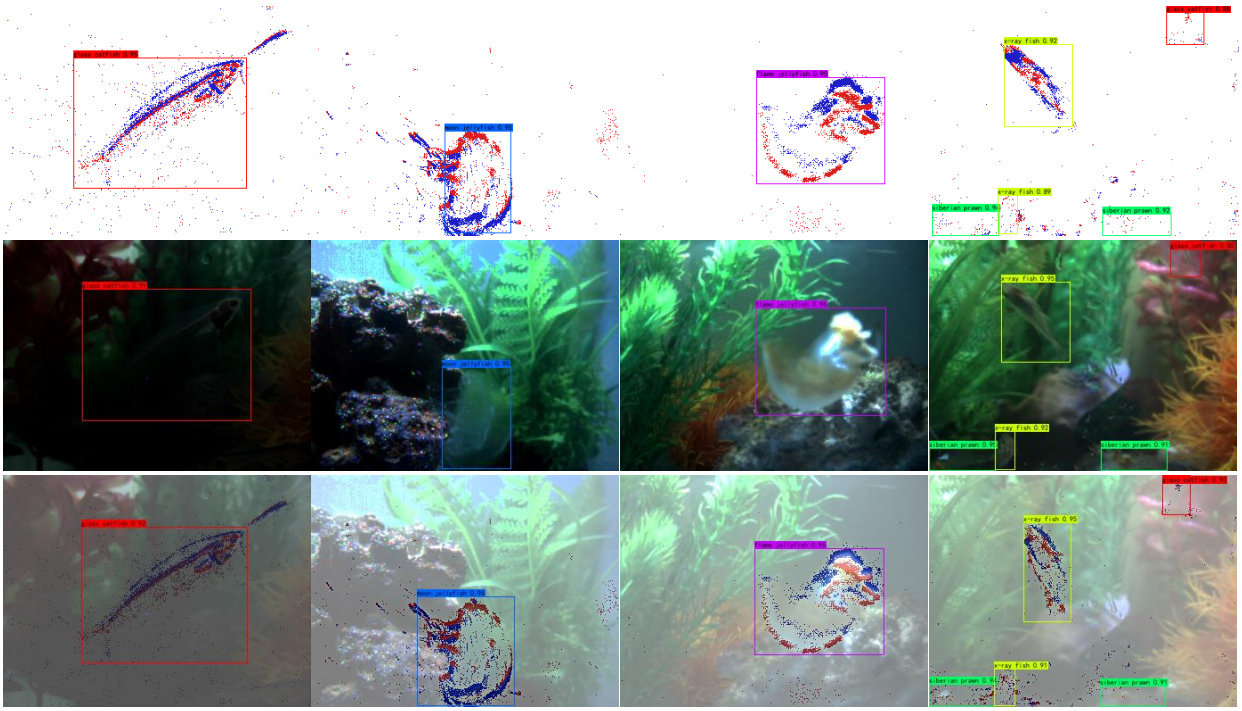


Fig. 8. Detection results of different image types in the dataset using TransCODNet. The top row displays the detection results of the event frames, the middle row displays the detection results of RGB, and the bottom line displays the detection results of the fused images.

TABLE II
AP AND MAP OF DIFFERENT ALGORITHMS

Image Type	Methods	AP(%)						mAP(%)
		Backbone	Flame Jellyfish	Glass Fish	Moon Jellyfish	Siberian Prawn	X-ray Fish	
Fused Image	CenterNet	Hourglass-104	82.3	80.2	82.5	81.1	79.6	81.1
	EfficientDet	Efficient-B0	85.7	80.2	82.4	74.8	76.0	79.8
	Faster R-CNN	ResNet50	85.1	77.0	78.9	75.5	72.2	77.7
	RetinaNet	ResNet50	86.0	78.9	81.2	76.8	77.2	80.0
	SSD	VGGNet	85.1	77.6	78.1	75.9	73.1	78.0
	Yolov3	DarkNet	81.1	71.0	72.1	67.4	65.2	71.4
	Yolov5	CSPDarkNet	87.4	84.1	84.5	82.0	78.2	83.2
	Yolov7	CBS+ELAN+MP	88.0	83.9	84.5	82.3	80.2	83.7
	YOLOX	CSPDarkNet	87.5	84.8	81.8	82.6	**81.2**	83.6
	TransCODNet	MobileDarkNet	**88.2**	**85.5**	**86.0**	**83.2**	80.7	**84.7**

TABLE III
ABLATION STUDY USING THE PRESENTED DATASETS, INCLUDING VERTICAL AND HORIZONTAL COMPARISONS

Image type	Method			mAP(%)	Params(M)	FLOPs(G)	FPS(f/s)
	MobileDarkNet	VEM	VarifocalLoss				
RGB				83.1	54.1	77.8	26.5
	✓			83.5	36.0	49.3	37.7
		✓		83.7	189.1	201.3	18.4
			✓	83.6	54.2	77.8	26.5
	✓	✓	✓	83.8	170.9	172.7	22.1
Event frame				72.1	54.1	77.8	26.5
	✓			73.0	36.0	49.3	37.7
		✓		74.1	189.1	201.3	18.4
			✓	73.6	54.2	77.8	26.5
	✓	✓	✓	74.5	170.9	172.7	22.1
Fused image				83.6	54.1	77.8	26.5
	✓			84.2	**36.0**	**49.3**	37.7
		✓		84.1	189.1	201.3	**18.4**
			✓	84.4	54.2	77.8	26.5
	✓	✓	✓	**84.7**	170.9	172.7	22.1

on the model with a fixed hyperparameter, and the image input size was 640×640. The ablation experiment results are shown in Table III, and '✓' represents the strategy used in this experiment. The types of images detected include

RGB images, event frames and fused images. This ablation experiment consists of two parts. The first part is to validate the effectiveness of the proposed method by using RGB, event frames, and fused images as inputs to observe the performance of the proposed improved algorithm under different input types. The second part is to use the same detection algorithm to separately detect RGB, event frames and fused images to verify the reliability of the fusion method used.

For the first part of the ablation experiment, improvement 1 is MoblieDarkNet, which represents changing the CSPNet part used in the backbone of YOLOX to MobileNet. Improvement 2 is VEM, which is proposed to expand the receptive field and enhance the transparent biological edge feature module, and improvement 3 is VarifocalLoss, which replaces BCELoss used in confidence loss in YOLOX to balance positive and negative samples. Table III demonstrates that for the detection fused images, the accuracy value of improvement 1 is 84.2% and the parameters and FLOPs are greatly reduced compared with the unimproved algorithm, and the FPS reaches 37.7f/s, which can be detected in real time. The accuracy of improvement 2 is increased by 0.5% compared with the unimproved algorithm. As VEM is a new module added on top of the original algorithm, its Parameters and FLOPs have both increased, resulting in a slower detection frame rate per second. The mAP of improvement 3 is 84.4%. Since only the loss function is changed, its parameters, FLOPs, and FPS are the same as those of the original algorithm. In the end, the accuracy value of TransCODNet is 84.7%, which is 1.1% higher than the unmodified YOLOX. The Parameters and FLOPs are slightly higher than YOLOX. Although lightweight MobileNet is used in TransCODNet, FPS does not meet the needs of real-time detection tasks due to the addition of VEM, which is also the direction of our next efforts. In the future, we will be committed to building a more lightweight real-time underwater transparent object detection framework.

For the second part of the ablation experiment, the accuracy value of YOLOX detection fused images is 83.6%, which is 11.5% higher than the detection of event frames and slightly higher than the detection accuracy of RGB. The algorithm for improvement 1, improvement 2, and improvement 3 has increased the mAP detection of fused images by 0.7%, 0.4%, and 0.8% compared to the mAP detection of RGB, respectively. The detection performance of event frames is lower than that of fused images and RGB. The accuracy of the TransCODNet algorithm in detecting fused images is 84.7%, which is 0.9% higher than the accuracy of detecting RGB and 10.2% higher than the accuracy of detecting event frames. Therefore, the model that combines RGB and event frames can better utilize color and texture information, thereby achieving better performance in these tasks.

In summary, this ablation experiment verifies the feasibility of the proposed three improved methods and also proves the effectiveness of pixel-level fusion of event frames and RGB in this letter.

VI. CONCLUSION

In this letter, we propose an underwater transparently camouflaged organisms dataset called Aqua-Eye, and a detection framework called TransCODNet. For this dataset, we use deep learning methods to fuse event frames with RGB to enhance the edge features of underwater transparently camouflaged organisms. At the same time, we have made three improvements based on YOLOX to improve detection accuracy. Firstly, we use MobileDarkNet to extract image features. Then, we propose a visual enhancement model(VEM) to further recognize the texture features of the object, and use VarifocalLoss to place the training focus on high-quality positive samples. The results showed that TransCODNet achieved the best results among the nine deep learning algorithms, ultimately achieving an accuracy of 84.7%.

We hope this work will raise awareness of the study of underwater transparently camouflaged objects. In the future, we will expand our dataset, add more sample types, and cover a wider range of application scenarios. In addition, in terms of algorithms, we will design a more lightweight saliency detection algorithm to achieve real-time detection of underwater transparently camouflaged organisms.

REFERENCES

[1] N.-E. HUSSEY, S.-T. Kessel et al. "Aquatic animal telemetry: a panoramic window into the underwater world," *Science*, 348(6240), 1255642,2015.

[2] E Laura, Bagge "Not as clear as it may appear: challenges associated with transparent camouflage in the ocean." *Integrative and Comparative Biology,* vol. 59, no. 6, pp.1653-1663, 2019.

[3] C. Brandli, R. Berner, M. Yang et al., "A 240× 180 130 db 3 µs latency global shutter spatiotemporal vision sensor," *IEEE J. Solid-State Circuits,* vol. 49, no. 10, pp. 2333-2341, 2014.

[4] E. Perot, P. de Tournemire, D. Nitti, J. Masci, and A. Sironi. "Learning to detect objects with a 1 megapixel event camera," *Adv. Neural Inf. Process. Syst.,* vol. 33, pp. 16639-16652, 2020.

[5] Z. Ge, S. Liu, F. Wang et al, "Yolox: Exceeding yolo series in 2021". *arXiv:2107.08430, 2021.*

[6] C.-Y. Wang, H.-Y.-M Liao, Y.-H. Wu et al., "CSPNet: A new backbone that can enhance learning capability of CNN," *in Proc. IEEE/CVF Int. Conf. Comput.* Vis., 2020, pp. 390-391.

[7] F. Zocco, T. C.Lin, C. I Huang et al., "Towards More Efficient Efficient-Dets and Real-Time Marine Debris Detection", *IEEE Robot. Autom. Lett.*, vol. 8, no. 4, pp. 2134-2141, 2023.

[8] F. Xu, H. Wang, J. Peng et al., "Scale-aware feature pyramid architecture for marine object detection," *Neural Comput. Appl.*, vol. 33, no. 8, pp. 3637–3653, 2021.

[9] M. Pedersen, R.Gade, T. B.Moeslund et al., "Detection of marine animals in a new underwater dataset with varying visibility", *in Proc. IEEE/CVF Int. Conf. Comput.* Vis., 2019, pp. 18-26.

[10] G. Gallego, T. Delbrück, G. Orchard et al., "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no.1, pp. 154-180. 2020.

[11] P. -F. Deng, G.-P Ji, G. Sun et al., "Camouflaged object detection," *in IEEE/CVF Int. Conf. Comput. Vis.,* 2020, pp. 2777-2787.

[12] H. Mei, G.-P Ji, Z. Wei et al., "Camouflaged object segmentation with distraction mining," *in Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8772–8781.

[13] J. Ren, X. Hu, L. Zhu et al., "Deep texture-aware features for camouflaged object detection". *IEEE Trans. Circuits Syst. Video Technol.*, 2021.

[14] K. Wang, H. Bi, Y. Zhang et al., "D2C-Net: A Dual-Branch, Dual-Guidance and Cross-Refine Network for Camouflaged Object Detection," *IEEE Trans. Ind. Electron.*, vol. 69, no.5, pp. 5364-5374, 2021.

[15] M. Pedersen, J. Bruslund Haurum, R. Gade et al., "Detection of marine animals in a new underwater dataset with varying visibility," *in Proc. IEEE/CVF Workshops Int. Conf. Comput.* Vis., 2019. pp. 18-26.

[16] Liu, Risheng et al., Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE Trans. Circuits Syst. Video Technol.*, 2020, 30.12: 4861-4875.

[17] Ch.-Y. Li et al., "An underwater image enhancement benchmark dataset and beyond," *IEEE Trans. Image Process.* vol. 29, pp. 4376-4389, 2019.

[18] X. Wang et al., "Visevent: Reliable object tracking via collaboration of frame and event flows," 2021, *arXiv:2108.05015.*

[19] H. Zhang, Y. Wang et al., "Varifocalnet: An iou-aware dense object detector". *in Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp.8514-8523.

[20] Loc Cordone, Benot Miramond, and Philippe Thierion. Object detection with spiking neural networks on automotive event data. *In Int. Joint Conf. Neural Netw. (IJCNN),* 2022.

[21] Zh.-y Jiang, P. Xia, K. Huang, "Mixed frame-/event-driven fast pedestrian detection," *in IEEE Int. Conf. Robot. Autom.*, 2019.

[22] H. Cao, G. Chen, J. Xia et al., "Fusion-based feature attention gate component for vehicle detection based on event camera," *IEEE Sens. J.*, vol. 21, no. 21, pp. 24540-24548,2021.

[23] A. Tomy, A. Paigwar, K. S. Mann et al., "Fusing event-based and RGB camera for robust object detection in adverse conditions," *in IEEE Int. Conf. Robot. Autom.*, 2022.

[24] H. Li, , X.-J. Wu, and K.Josef. "Infrared and visible image fusion using a deep learning framework," *in Int. Conf. Pattern Recognit*., 2018.

[25] Kugele, Alexander, et al., "Hybrid SNN-ANN: Energy-efficient classification and object detection for event-based vision." *DAGM German Conference on Pattern Recognition.* Cham: Springer International Publishing, 2021.

[26] S. Liu, D. Huang. "Receptive field block net for accurate and fast object detection," *in Proc. Eur. Conf. Comput.* Vis. 2018, pp.385-400.

[27] X. Zhou, D. Wang and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850.*

[28] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutionXal neural networks," *in Proc. Int. Conf. Mach. Learn*., pp. 6105–6114, 2019.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *in Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[31] W. Liu et al., "SSD: Single shot multibox detector," *in Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[32] J. Redmon, A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767.*

[33] C.-Y Wang, A. Bochkovskiy et al., "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696.*

[34] C. Yang, Z. Zhang et al., "FakeMix augmentation improves transparent object detection," 2021, arXiv:2103.13279.

[35] H.-Y. Mei, X. Yang, Y. Wang et al., "Don't hit me! glass detection in real-world scenes," in Proc. IEEE/CVF Int. Conf. Comput. Vis. 2020, pp. 3687-3696.

[36] J.-Q. Jiang, G.-Q. Cao, T.-T. Do et al., "A4T: Hierarchical affordance detection for transparent objects depth reconstruction and manipulation", IEEE Robot. Autom. Lett., vol. 7, no. 8, pp. 9826-9833, 2020.