Deep Learning-Based Cascaded Light Source Detection for Link Alignment in Underwater Wireless Optical Communication

Bowen Jia, Wenmin Ge, Jingxuan Cheng, Zihao Du, Renming Wang, Guangbin Song, Yufan Zhang, Chengye Cai, Sitong Qin, and Jing Xu

***Abstract*—Obtaining the light source position from the image is an important solution for achieving link alignment in laser-based underwater wireless optical communication (UWOC) systems. However, in practical scenarios, the misalignment degree between the light source and camera is variable, and factors such as ambient light may introduce disturbances, leading to significant variations in the appearance of light spots in images. Existing research primarily relies on simple features like brightness, color, or shape, which makes it difficult to accurately obtain position information from these non-ideal images. In this paper, deep neural networks (DNNs) with strong feature extraction capabilities are introduced to automatically learn the patterns of the light source from diverse images. A detection architecture cascading an object detector and a keypoint detector is adopted, achieving better comprehensive performance in terms of accuracy and speed. To train and evaluate the deep learning model, we construct the UWOC Light Source Detection Benchmark (ULDB) dataset. This dataset comprises 2200 images captured in a standard swimming pool, covering a misalignment range far beyond existing studies. On the ULDB test set, the proposed detection method achieves an average precision (AP) of 99.1% and an average positioning error of 4.66 pixels, while the traditional method may frequently extract false light spots. To the best of our knowledge, the ULDB dataset is the first image dataset specifically designed for the task of link alignment between UWOC terminals.**

***Index Terms*—Communication link alignment, deep learning, object detection, keypoint detection, underwater wireless optical communication (UWOC).**

Received 17 August 2024; accepted 28 August 2024. Date of publication 2 September 2024; date of current version 13 September 2024. This work was supported in part by Key Research and Development Program of Hainan Province under Grant ZDYF2023GXJS016, in part by the Project of Sanya Yazhou Bay Science and Technology City under Grant SCKJ-JYRC-2022-40, and in part by National Key Research and Development Program of China under Grant 2022YFC2808200, Grant 2022YFB2903403, and Grant 2022YFC2808100. *(Corresponding author: Jing Xu.)*

Bowen Jia, Renming Wang, Guangbin Song, and Jing Xu are with the Hainan Institute of Zhejiang University, Sanya 572025, China, also with the Optical Communications Laboratory, Ocean College, Zhejiang University, Zhoushan 316021, China, and also with the ZTT-Ocean College Joint Research Center for Marine Optoelectronic Technology, Ocean College, Zhejiang University, Zhoushan 316021, China (e-mail: jybw@zju.edu.cn; wrm@zju.edu.cn; 12034094@zju.edu.cn; jxu-optics@zju.edu.cn).

Wenmin Ge, Jingxuan Cheng, Zihao Du, Yufan Zhang, Chengye Cai, and Sitong Qin are with the Optical Communications Laboratory, Ocean College, Zhejiang University, Zhoushan 316021, China, and also with the ZTT-Ocean College Joint Research Center for Marine Optoelectronic Technology, Ocean College, Zhejiang University, Zhoushan 316021, China (e-mail: 12034062@zju.edu.cn; 12134076@zju.edu.cn; 11934056@zju.edu.cn; 12034050@zju.edu.cn; choi122304057@zju.edu.cn; sitonqin@zju.edu.cn).

The dataset is available at https://github.com/happycode123/ULDB-Dataset.

Digital Object Identifier 10.1109/JPHOT.2024.3453116

I. Introduction

EFFICIENT underwater communication systems play an important role in interconnectivity of future Internet of Underwater Things (IoUT) devices. As one of the means of underwater communication, underwater wireless optical communication (UWOC) has attracted wide attention due to its high bandwidth, high data rate, low power consumption, and appropriate transmission distance [1], [2]. Laser diodes (LDs) and light-emitting diodes (LEDs) are commonly used as light sources for UWOC. Compared to LEDs, LDs can provide higher modulation bandwidth and better collimation. Many UWOC systems have greatly benefited from the excellent characteristics of LDs in terms of transmission distance and data rate [3], [4], [5], [6], [7].

Despite these strengths of LD-based UWOC systems, their strict alignment requirements cannot be ignored in practical applications. Existing solutions can be divided into passive methods and active methods. Passive methods aim to alleviate the alignment requirement by enhancing the propagation capability of the transmitter or the sensing ability of the receiver [8], [9], [10], [11]. In contrast, active methods make the UWOC terminals actively point at each other by moving or rotating, and then the line-of-sight (LOS) link is established [12], [13], [14], [15], [16]. For active methods, it is essential to obtain accurate position information about the light source during the alignment process.

Many UWOC active alignment (UAA) systems use the camera to search for the light source and obtain its position from the captured image using positioning algorithms. In [17], grayscale centroid method was employed to extract the light source position from images. Tracking algorithms proposed in [19] and [23] can obtain the light spot position from consecutive video frames. To enhance the alignment capability of camera-assisted UAA systems, Williams et al. adopted a large field-of-view (FOV) camera, expanding the tolerable range of camera deviation angle (a detailed definition is presented in Section II) [17]. If the light source is not pointed at the camera, the shape of the spot in the image becomes distorted. In view of this, AprilTag detection was

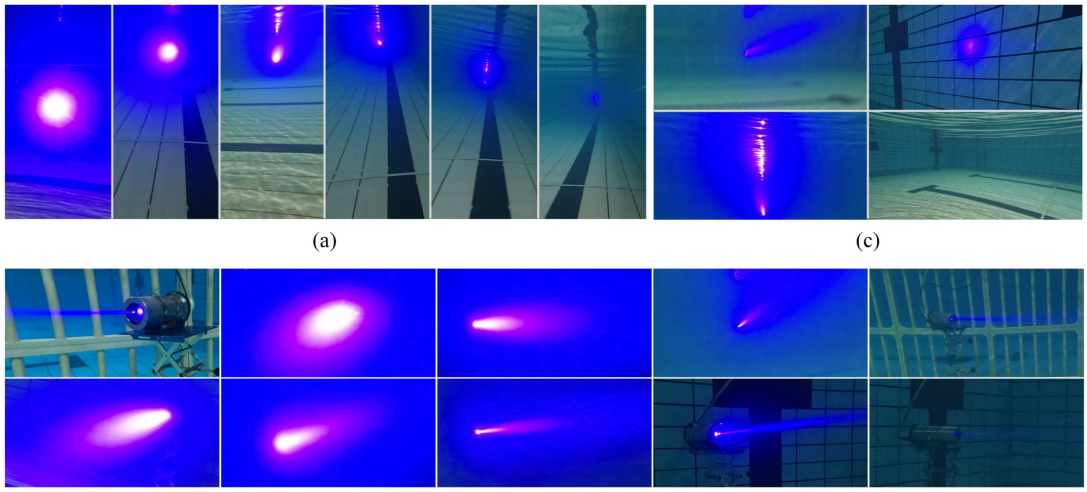

Fig. 1. Challenges in light source detection. (a) Variation in spot size. (b) Variation in spot shape. (c) Other challenges such as motion blur, surface reflection, pure background.

TABLE I
COMPARISON OF CAMERA-ASSISTED UAA SYSTEMS

Reference	FOV of Camera	Positioning Method	Tolerable Terminal-camera Distance Range	Tolerable Terminal Deviation Angle (absolute value) Range
[17]	100°	GT[a] + Centroid	No more than 7.5 m	Small
[18]	3.6°	/	20 m ~ 25 m	Small
[19]	/	MASMS[b]	/	/
[20]	/	AprilTag detection	No more than 6 m	Medium (0° ~ 60°)
[21]	/	GT + Spot shape analysis	About 16 m	Small (0° ~ 8°)
[22]	/	GT + PCA[c]	About 15 m	Small (1° ~ 5°)
[23]	/	Camshift + Kalman	2 m	/
This work	87°	Deep learning-based cascaded light source detection	**1 m ~ 32 m**	**Large** (0° ~ nearly 180°)

[a] GT represents gray threshold.
[b] MASMS represents modified adapted scale mean-shift.
[c] PCA represents principal component analysis.

introduced into the camera-assisted UAA system, significantly expanding the tolerable range of terminal deviation angle [20]. Tang et al. utilized the relationship between receiver deviation distance and spot shape to achieve alignment [21]. A positioning method based on principal component analysis (PCA) proposed in [22] can extract the light source position from distorted light spots.

The aforementioned methods work well when the terminal-camera misalignment degree is relatively small and varies slightly. In practice, however, the terminal-camera misalignment degree is normally arbitrary before the alignment process. This implies that the terminal-camera distance varies largely, while the deviation angles of the terminal or camera also range from −180 degrees to 180 degrees. As shown in Fig. 1, with the change of misalignment degree, the light spot in the image may undergo a significant variation in terms of position, size, shape, and brightness. Interference such as reflection is also common. Existing positioning methods mainly rely on simple features like brightness, color, or shape, making it challenging to accurately extract the position information of light source from these non-ideal images. Therefore, the misalignment ranges that these methods can handle are limited (see Table I).

Inspired by the success of deep learning in advanced computer vision tasks such as classification, detection, and segmentation [24], [25], [26], [27], [28], we introduce deep neural networks (DNNs) to obtain accurate light source position from captured images with various light spot sizes and shapes. According to the knowledge from the computer vision community, this positioning task is defined as UWOC light source detection. Specifically, UWOC light source detection refers to determining whether a communication light source exists in an image. If it exists, its position in the image is further provided; if it does not exist, a lower prediction score is given to indicate this result. To achieve better comprehensive performance in accuracy and

speed, we consider light source detection as a cascaded combination of object detection and keypoint detection. Thanks to the powerful feature extraction capability of the DNN, the proposed method shows significant improvement compared to the existing works in terms of tolerable range of misalignment. Its tolerable terminal-camera distance ranges from 1 to 32 meters, while its tolerable terminal deviation angle (absolute value) ranges from 0 to nearly 180 degrees. When facing complex cases that may make the existing methods fail, such as water surface reflection, background light interference, and more severe spot distortion, the proposed method can still provide correct detection results. For the 4K images, the proposed method achieves an average precision (AP) of 99.1% and an average positioning error of 4.66 pixels. It is worth noting that this method requires no human intervention, that is, the original image is directly input into the detector without manually extracting any potential regions of interest.

To train and evaluate deep learning models, we also construct the UWOC Light Source Detection Benchmark (ULDB) dataset, which consists of 2200 images with annotated positions. These images were all captured using a camera with a large FOV of 87 degrees in a standard swimming pool. They cover a wide range of misalignment scenarios, including diverse variations in distance, angle, and external interference factors. To the best of our knowledge, this is the first image dataset specifically designed for the task of link alignment in the UWOC field. It can be used not only for training data-driven models but also as a fair platform to evaluate the performance of light source detectors.

The rest of this paper is organized as follows: Section II introduces the description of misalignment degree and discusses the effect of misalignment degree on imaging results. In Section III, the details about ULDB dataset are illustrated, and the deep learning-based light source detection method is proposed. In Section IV, the experimental results are presented and discussed. Finally, our work is summarized in Section V.

II. Description of Misalignment Degree

In this paper, we define the deviation angle to clearly and quantitatively describe the misalignment degree. With the help of some devices such as pressure sensors, it is not difficult for the underwater terminal to reach a specific depth [18], so we only consider the deviation angle in the horizontal plane.

As shown in Fig. 2(a), the deviation angle θ_d of a terminal is defined as

$$\theta_d = \begin{cases} \theta_{CW}, & |\theta_{CW}| \le |\theta_{CCW}| \\ \theta_{CCW}, & |\theta_{CW}| > |\theta_{CCW}| \end{cases}, \tag{1}$$

where θ_{CW} and θ_{CCW} represent the angle required for the terminal to rotate clockwise and counterclockwise, respectively, from its current orientation to the aligned orientation. It is specified that negative values denote clockwise rotation, while positive ones denote counterclockwise rotation. The range of θ_d is $(-180°, +180°]$. Fig. 2(b) shows some examples about the deviation angle. It can be observed that the respective deviation angles of Terminal A and Terminal B explicitly reflect alignment/misalignment status between them. When both of their deviation angles are 0°, the whole UWOC system achieves alignment. In contrast, as long as the deviation angle of either terminal is not 0°, it cannot be considered that the UWOC system is aligned. For two terminals in a plane, the combination of their individual deviation angles and the distance between them can clearly describe all misalignment scenarios between them.

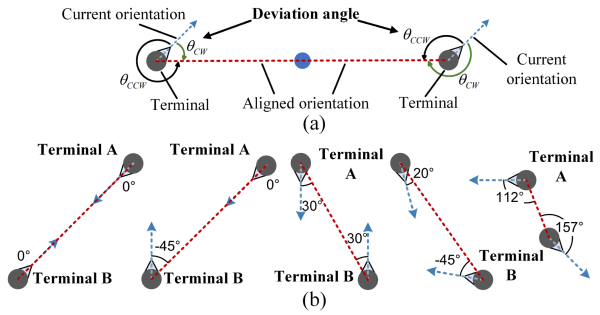

Fig. 2. (a) Definition of deviation angle. (b) Some examples about deviation angle. The orientation of each terminal is indicated by an arrow and a triangle.

After defining the deviation angle, its effect on camera-assisted UAA systems is examined in this paper. Fig. 3 shows the imaging results at different deviation angles. As the $|\theta_d|$ of the terminal increases, the light spot shape changes from a perfect circle to a beam-like shape, and the brightness of the light spot gradually diminishes. As the $|\theta_d|$ of the camera increases, the light spot gradually moves from the center of the image towards the edge until it moves out of the image. In addition, as shown in Fig. 1(a), the size and brightness of the light spot decrease as the camera moves away from the terminal. In summary, the deviation angle of the terminal affects the shape and brightness of the light spot, the deviation angle of the camera affects the position of the light spot, and the terminal-camera distance affects the size and brightness of the light spot.

The deviation angle of the terminal also has an effect on the relative position relationship between the light source and the light spot. In this paper, the position of the light source is represented as a point (i.e., coordinate) in the image for alignment purposes. When the terminal deviation angle is close to 0°, the light spot appears as an ideal circle (Fig. 4(a)). It is generally regarded that the center of the light spot is the light source position [21]. When the $|\theta_d|$ of the terminal reaches a certain level, the light spot distorts into an obvious beam shape (Fig. 4(e) and (f)). At this point, the emission port of the light source on the terminal is exposed and the light source is located at the edge of the beam. Therefore, as the light spot gradually distorts from a circular shape into a beam shape, the light source position gradually moves from the center of the light spot to the edge. Fig. 4 illustrates this process.

III. Method

According to the discussion in Section II, the position, size, brightness, and shape of light spots in images are all variables. Therefore, extracting light spots based on simple image features is likely to be inaccurate. Moreover, it is difficult to explicitly model the relative position relationship between the

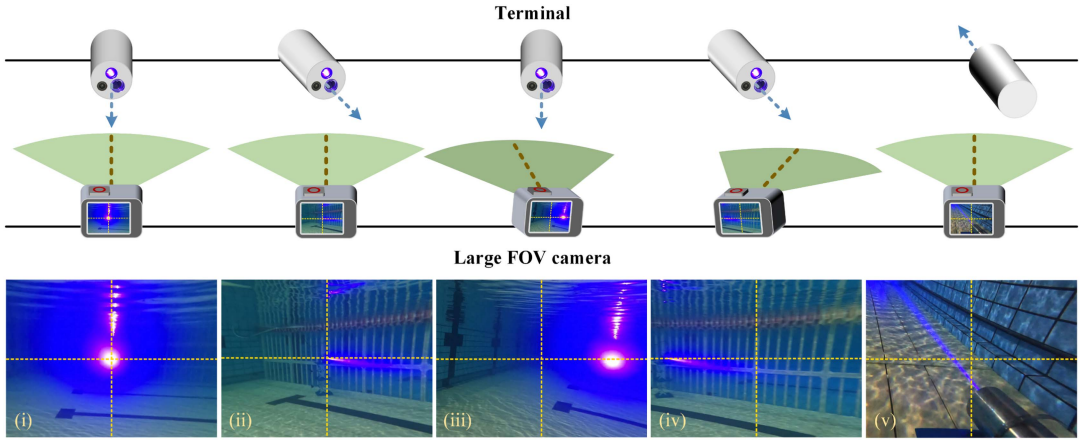

Fig. 3. Effect of the deviation angle on the imaging result. Insets: (i) When the terminal and camera are both aligned with each other, the light spot appears round and is located in the center of the captured image. (ii)∼(v) If the terminal faces other directions, the spot shape changes. If the camera faces other directions, the spot position changes.

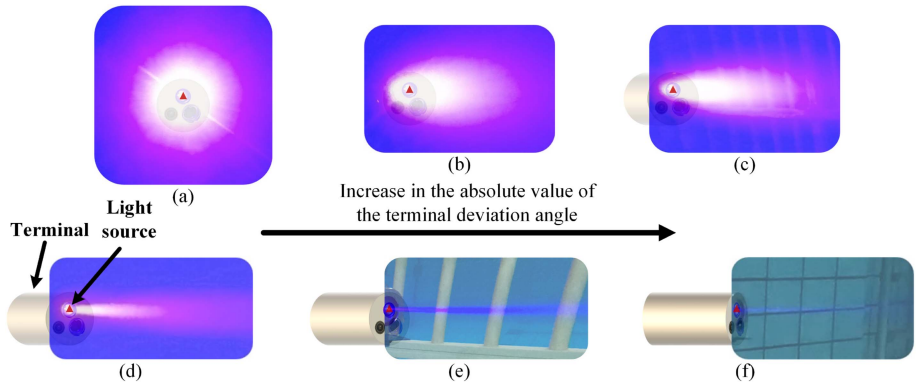

Fig. 4. Effect of the terminal deviation angle on the relative position relationship between the light source and the light spot. From (a) to (f), the absolute value of the terminal deviation angle gradually increases. The light source position is marked by a red triangle.

light source and the light spot. To extend the application of the camera-assisted UAA system in more general alignment scenarios, the DNN is introduced to detect the light source from images.

A. ULDB Dataset

Deep learning-based methods require high-quality datasets for training and evaluation, so we constructed the ULDB dataset. We used a 450-nm blue LD with a front collimating lens as the light source. The output optical power of the LD was about 23.78 dBm (238.8 mW). The entire light source was fixed in a watertight cabin, serving as the UWOC terminal. We used a GoPro HERO10 Black waterproof camera to capture light source images with linear mode, which provided a horizontal FOV of 87°. The resolution and frame rate were set as 3840×2160 and 60 fps, respectively.

As shown in Fig. 5, we demarcated a 35 m × 11 m rectangular area in an indoor standard swimming pool for data collection. The terminal was submerged at a depth of 0.6 m. By measuring the received optical power at different distances, the attenuation coefficient of the pool water was estimated to be approximately 0.092 m^{-1}. During data collection, the position and orientation of the terminal were adjusted multiple times, while the position and orientation of the camera were constantly changing. The ambient light condition was also changed, since we collected image data under different scenarios, including "nighttime with indoor lights on/off" and "daytime with indoor lights on/off". In addition, no zoom-in was employed, implying that the FOV of the camera was always 87°.

After data collection, 22673 images were taken from the captured videos at an interval of 10 frames. From these images, we further carefully selected the 2200 most representative images as the ULDB dataset. Among these 2200 images, 300 are pure background images without any terminal, which can help the detection model to reduce "false-positive" errors. Among the remaining 1900 images with the terminal, the terminal-camera distance ranges from 1 m to 32 m, the $|\theta_d|$ of the terminal varies from 0° to nearly 180° (i.e., behind the terminal), and the maximum $|\theta_d|$ of the camera is more than 45° (i.e., the light source was outside the FOV, but a part of the light spot appeared at the edge of the captured image). Moreover, a large number of images with motion blur or reflection interference are retained in the ULDB dataset to reflect the real working

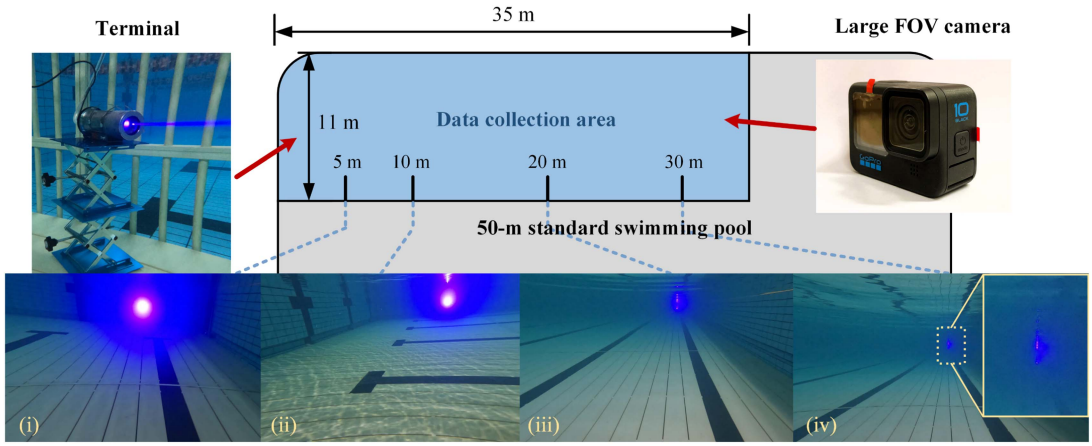

Fig. 5. Data collection scenario. Insets: Typical image captured at a distance of (i) 5-m, (ii) 10-m, (iii) 20-m, and (iv) 30-m from the terminal.

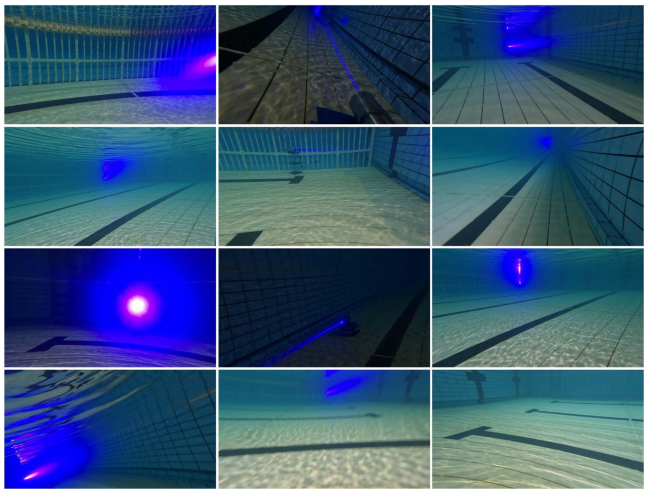

Fig. 6. Examples of the images in ULDB dataset.

environment. Overall, the ULDB dataset exhibits good diversity. Some examples of images in this dataset are presented in Fig. 6.

According to a certain ratio, the ULDB dataset is also randomly divided into training/validation/test set. These subsets are respectively used for training the deep learning model, monitoring the training process, and evaluating the model performance. Some detailed information about the ULDB dataset is provided in Table II.

TABLE II
DETAILS OF THE PROPOSED ULDB DATASET

Type	Number
Total images	2200
Background images	300
Images in training/validation/test set	1540/330/330
Light spot region annotations	1900
Light source annotations	1849

B. Cascaded Light Source Detection Based on Deep Learning

In the computer vision community, detecting specific regions from an image is regarded as the object detection task, while detecting specific points from an image is known as the keypoint detection task. Therefore, light source detection is a keypoint detection task. However, it is inefficient to directly perform keypoint detection on the 4K images in the ULDB dataset. In order to speed up the detection, the high-resolution original image is commonly resized to a low-sized one before being processed by the DNN. But this process results in a loss of local features of the image, which reduces the accuracy of keypoint localization. Considering the above factors, we adopt a cascade detection architecture to achieve better comprehensive performance in terms of accuracy and speed.

As shown in Fig. 7, the original image is first fed into a deep learning-based object detection model, which gives the bounding box of the light spot region. Subsequently, the light spot region in the original image is cropped out as the input of a deep learning-based keypoint detector that eventually predicts the light source position. In this architecture, the light spot region detected by the object detection model helps the keypoint detector exclude a large number of irrelevant regions. Although resizing the original image is still unavoidable, more image details that benefit keypoint detection can be retained and are fed into the DNN of the keypoint detector.

As one of the latest detection models, YOLOv8 [29] and RTMPose [30] have achieved excellent performance in the fields of general object detection and human keypoint detection, respectively. More importantly, they are lightweight and easy to deploy, which is very important for underwater devices with limited power consumption. Therefore, they are used to construct the cascaded light source detector in our work. Fig. 7 also shows the internal architecture of YOLOv8 and RTMPose. Generally speaking, they include five modules: pre-processing, backbone, neck, head, and post-processing. The pre-processing module adjusts the original image to an input tensor suitable for DNN

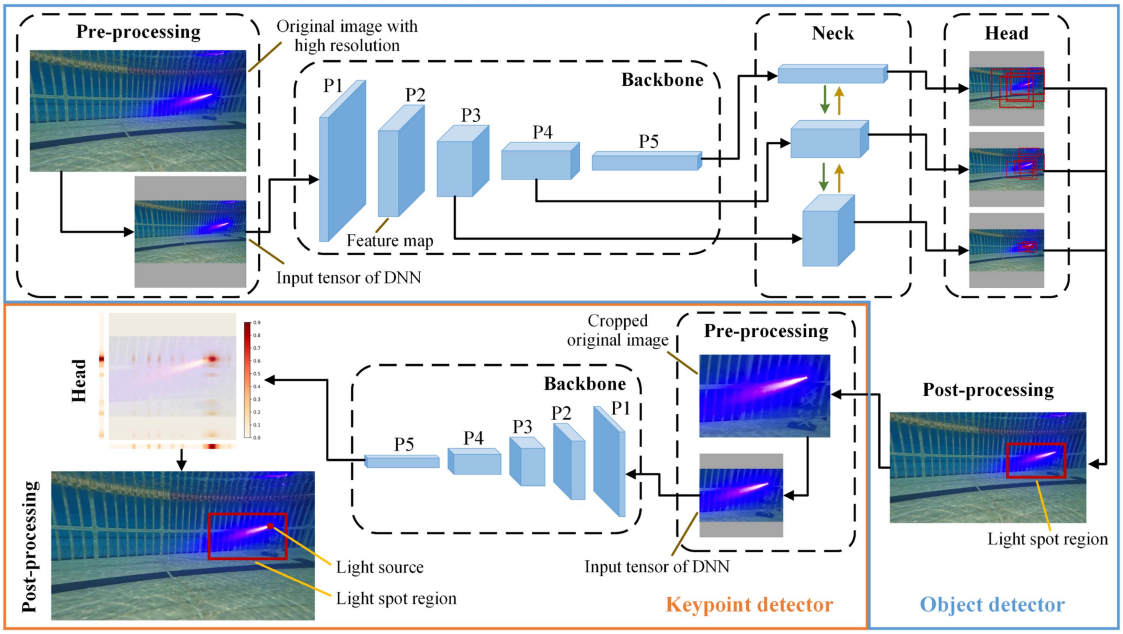

Fig. 7. Architecture of the proposed cascaded light source detector based on deep learning. Backbone module is composed of deep CNN and is responsible for feature extraction. P1∼P5 in backbone represent feature maps at different levels, which have different resolutions and characterize semantic information at different levels. P1 has the highest resolution and the lowest-level semantic information, while P5 has the lowest resolution and the highest-level semantic information. Details of other modules are provided in the text.

processing in terms of size and numerical range. The backbone is the main part of the detector and is used to extract image features. As the input tensor is forward propagated in the backbone, feature maps with different levels and sizes are sequentially generated. The feature maps from P1 to P5 are downsampled relative to the input tensor by factors of 2×, 4×, 8×, 16×, and 32×, respectively. The neck module is responsible for fusing the feature maps of P3, P4, and P5. The head module obtains preliminary information about the object location and category based on the feature maps processed by the neck module. The post-processing module decodes and filters the results provided by the head, and then outputs the final prediction results. In these five modules, the backbone, neck and head modules are typically composed of convolutional neural networks (CNNs). The backbone to head forms an end-to-end DNN. It is noted that the head module of YOLOv8 outputs three prediction tensors with different sizes, which collectively contribute to detecting light spot regions. This design is conducive to the model to cope with the variation of spot size. In contrast, RTMPose adopts a more lightweight design without a neck module. Its head module obtains preliminary information about the light source position directly from the P5 feature map output by the backbone.

Unlike traditional methods, deep learning models do not require the handcrafted feature representations. Through training, the model can automatically learn the potential features of the target based on image-annotation pairs. During training, the output tensors of the head module are compared with the same-sized tensors generated from annotations to compute the loss. This loss is then used to optimize the parameters of the DNN so that the model can ultimately produce the desired results. After being trained, when the original image is input into YOLOv8, the center coordinates, width, height, and predicted scores of the light spot regions will be output. Similarly, when the cropped image is input into RTMPose, the coordinates and predicted scores of the light source will be output. Both YOLOv8 and RTMPose ultimately output position predictions based on the scale of the original image.

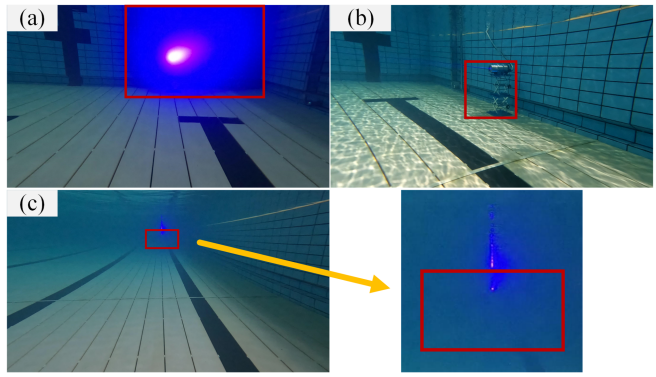

Fig. 8. Illustration of the annotation strategy of the light spot region.

In our proposed detection method, how to define the light spot region is also a critical issue. A strict definition makes it difficult to annotate images and predict bounding boxes, while a loose definition results in too many invalid regions in the bounding box and weakens the keypoint detector performance. To obtain a reasonable light spot region, four annotation strategies are adopted:

1) The halo outside the spot is considered as part of the light spot region (Fig. 8(a)).

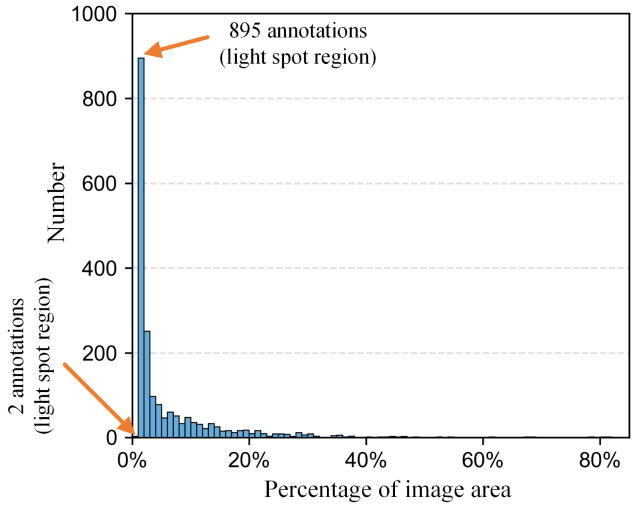

Fig. 9. Area distribution of the annotated bounding boxes of the light spot region in the ULDB dataset. The area of each bounding box is represented as a percentage of the image area.

2) The image area occupied by the terminal is also considered as part of the light spot region, if its appearance in the image is clearly recognizable (Fig. 8(b)).
3) For the images with very small light spot region, instead of annotating the bounding box just fitting the edges of the light spot region, a moderate-size bounding box is annotated around the light source (Fig. 8(c)). In other words, the bounding box can include some background in case that it is too small.
4) When annotating the light spot region, all reflections are carefully avoided to train the object detection model to exclude the reflection interference (Fig. 8(c)).

Fig. 9 shows the area distribution of the 1900 annotated light spot region bounding boxes in the ULDB dataset. The area is expressed as a percentage of the corresponding original image area. Only two bounding boxes have an area less than 1% of the corresponding image area. With respect to the light source keypoint, we follow the relative position relationship discussed in Section II for annotation.

IV. Experiment Results and Discussion

Both the training and testing procedures were performed on a desktop computer with Intel Core i7-13700F CPU and NVIDIA GeForce RTX 3060 GPU. The object detector and the keypoint detector were all trained from scratch. Mosaic data augmentation was used during the training of the object detector but was closed for the last 20 epochs. Other training hyperparameters are shown in Table III.

A. Evaluation Metrics

The AP metric and the percentage of correct keypoints (PCK) [31] are used to evaluate the detection performance. The AP metric can comprehensively reflect the precision and recall of a model. It can be considered as the area under the precision-recall curve. The precision P and recall R can be expressed as

$$P = \frac{C}{M}, \tag{2}$$

$$R = \frac{C}{N}. \tag{3}$$

In the above equations, C represents the number of correct predictions generated by the model, M represents the total number of predictions output by the model, and N represents the total number of annotations. Intersection over Union (IoU) and object keypoint similarity (OKS) are used to assess whether the object detection results and keypoint detection results are correct, respectively. In this paper, AP50 denotes the AP score at an evaluation threshold of 0.50, while AP denotes the average of AP scores at ten evaluation thresholds (i.e., 0.50, 0.55, …, 0.90, 0.95). A higher AP means that the model has a better discrimination between the object and the background. When calculating AP for a keypoint detector, a constant σ is introduced to balance the fluctuation arising from manual annotation. A smaller σ indicates lighter fluctuation and a stricter AP metric. We refer to the standard setting in the human keypoint detection task (i.e., 0.025 for the eyes, 0.087 for the knees), and set the σ of light source to 0.036. For more details about the AP metric, we refer readers to the MS COCO websites [32], [33].

TABLE III
Training Hyperparameter Configuration of YOLOv8 and RTMPose

Hyperparameters	YOLOv8	RTMPose
Epochs	300	150
Batch size	32	32
Base learning rate	0.01	0.001
Optimizer	SGD	AdamW
Warm-up epochs	3	5

The PCK metric examines whether the keypoint is correctly detected from another perspective. If the Euclidean distance between the predicted keypoint and its corresponding ground truth is less than the distance threshold d_{thr}, the detection is considered successful. the distance threshold d_{thr} can be written as

$$d_{thr} = \alpha \times \max(h, w). \tag{4}$$

In the above expression, h and w represent the height and width of the bounding box where the keypoint is located, and α is a hyperparameter. A smaller α indicates a stricter PCK metric. Common values for α are 0.1 and 0.2, but in our work, a more rigorous value is adopted. PCK under a specific α is denoted as PCK@α. As the terminal-camera distance increases, the size of light spot region decreases rapidly. Thus, the farther the light source is from the camera, the more accurate prediction is required to achieve the same PCK score.

In addition, the number of parameters and floating-point operations (FLOPs) of a deep learning model can largely characterize its storage and computation overhead, which determine whether the model can eventually be deployed to low-power underwater terminals. Therefore, the two metrics are also reported.

TABLE IV
PERFORMANCE OF YOLOV8 OBJECT DETECTION MODEL ON THE ULDB TEST SET

Model	Input Size	AP	AP50	Parameters	FLOPs
YOLOv8n	320×320	77.2	98.8	3M	2G
	416×416	78.8	**99.0**		3.5G
YOLOv8s	320×320	**80.2**	**99.0**	11.1M	7.2G
	416×416	80.0	**99.0**		12.1G

B. Quantitative Evaluation

By adjusting the number of layers and feature map channels in backbone, YOLOv8 and RTMPose with different scales can be generated. To determine the appropriate model scale, we first evaluated the detection performance of YOLOv8 for the light spot region on the ULDB test set (see Table IV). To our surprise, YOLOv8n, the smallest version of YOLOv8, shows satisfactory detection performance even if the resolution of the original image is as high as the 4K standard and most light spot regions are indeed very small. YOLOv8n is advantageous for underwater devices due to its lightweight design. When the input size of YOLOv8n is set as 320×320, its FLOPs are only 2G, while maintaining a high AP50 of 98.8%. To ensure better results for the subsequent keypoint detection, we select YOLOv8s with 320×320 input size as the object detector considering the trade-off between model accuracy and computational effort.

Next, we performed the full two-stage light source detection process on the test set. Thanks to the refinement function of the object detector, the input size of RTMPose does not need to be too large. We set it to 256×256, which effectively reduces the computation of the model. The results in Table V demonstrate that RTMPose-t (i.e., the smallest version of RTMPose) combined with YOLOv8s achieves the best performance on all evaluation metrics except for PCK@0.005. The accuracy-computation trade-off makes us select YOLOv8s 320×320 + RTMPose-t 256×256 as our final light source detector.

For the 330 images in the ULDB test set, the average positioning error of this method is only 4.66 pixels. When α is set as 0.03, the PCK reaches 100%, indicating that this method detects all light sources in the test set with considerable accuracy. Since the ULDB dataset covers abundant and various misalignment cases, these results demonstrate that the range of misalignment our method can handle is significantly expanded. For the scenes with a terminal-camera distance ranging from 1 m to 32 m and a terminal deviation angle (absolute value) ranging from 0° to nearly 180°, our method can provide sufficiently accurate predictions of light source positions.

Our method is also lightweight enough as its FLOPs are 8.1G, and its inference time on GPU is only 9.1 ms. It is worth noting that the inference speed was tested without any skip-frame detection, quantization, or deployment techniques, all of which can further accelerate the inference process. This suggests that the proposed method has the potential to be executed in real time on low-power underwater devices.

Fig. 10 illustrates the training process of YOLOv8s and RTMPose-t, respectively. In the early stage of the training process, AP50 of both models quickly approach the upper bound of 100%. As the training progresses, the scores on other metrics (more stringent) continuously improve. On the one hand, this indicates that the proposed method is easy to converge. On the other hand, it also demonstrates that the ULDB dataset is well representative and enables continuous improvement of the detection capability of the deep learning models without overfitting.

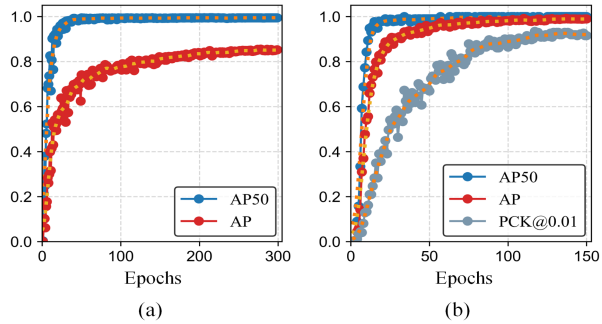

Fig. 10. Training process of (a) YOLOv8s and (b) RTMPose-t. The performance was tested on the ULDB validation set.

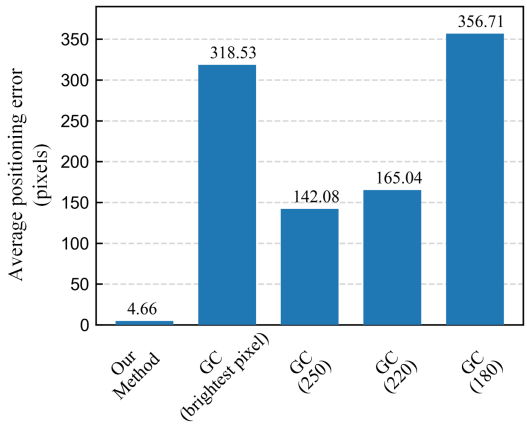

Fig. 11. Comparison with the grayscale centroid method on the ULDB test set. GC represents the grayscale centroid method. Values in parentheses indicate the thresholds used for extracting the light spot.

We also evaluated the performance of two single-stage detection models. As these models lack a coarse-to-fine refinement process, we set their input size to 640×640 to help them perceive as many image details as possible. By adding a keypoint detection head after the neck module of YOLOv8, YOLOv8-pose can simultaneously output the results of object detection and keypoint detection in one inference [34]. This scheme allows the two detection tasks to share a set of feature representations, and thus, high-precision keypoint localization is difficult to take into account. For instance, the YOLOv8m-pose has 81.2G FLOPs, but its AP is less than half that of our method, which does not meet the practical alignment requirements. The single-stage RTMPose directly detects the light source based on the original image. Although the RTMPose-t (single) has a significant advantage in computational complexity, its PCK@0.005 and PCK@0.01 decrease by 56.5% and 26.7%, respectively, compared with our

TABLE V
COMPARISON WITH OTHER DEEP LEARNING-BASED LIGHT SOURCE DETECTION METHODS ON THE ULDB TEST SET

Method	Input Size[a]	AP	AP50	PCK@				Params	FLOPs
				.005	.01	.02	.05		
YOLOv8s + RTMPose-t	256×256	**99.1**	**100.0**	67.2	**92.1**	**98.6**	**100.0**	**14.2M**	**8.1G**
YOLOv8s + RTMPose-s	256×256	98.4	**100.0**	**68.6**	91.3	98.2	99.6	16.2M	8.9G
YOLOv8n-pose	640×640	39.2	70.9	2.9	10.8	33.9	84.1	**3.1M**	**8.4G**
YOLOv8s-pose	640×640	46.4	76.5	4.3	16.3	**46.2**	85.9	11.4M	29.6G
YOLOv8m-pose	640×640	**48.8**	**79.8**	**6.1**	**20.2**	44.8	**86.6**	26.4M	81.2G
RTMPose-t (Single)	640×640	92.7	97.4	29.2	67.5	95.3	98.6	**6.8M**	**5.5G**
RTMPose-s (Single)	640×640	94.6	97.7	32.1	74.0	96.4	98.9	8.8M	10.8G
RTMPose-m (Single)	640×640	96.1	**98.9**	44.0	82.7	97.5	**99.6**	16.7M	31.3G
RTMPose-l (Single)	640×640	**97.6**	98.8	**50.9**	**87.7**	**98.2**	**99.6**	30.6M	68.2G

[a] The input size of YOLOv8+RTMPose method represents the input size of RTMPose. The input size of YOLOv8 is 320×320.

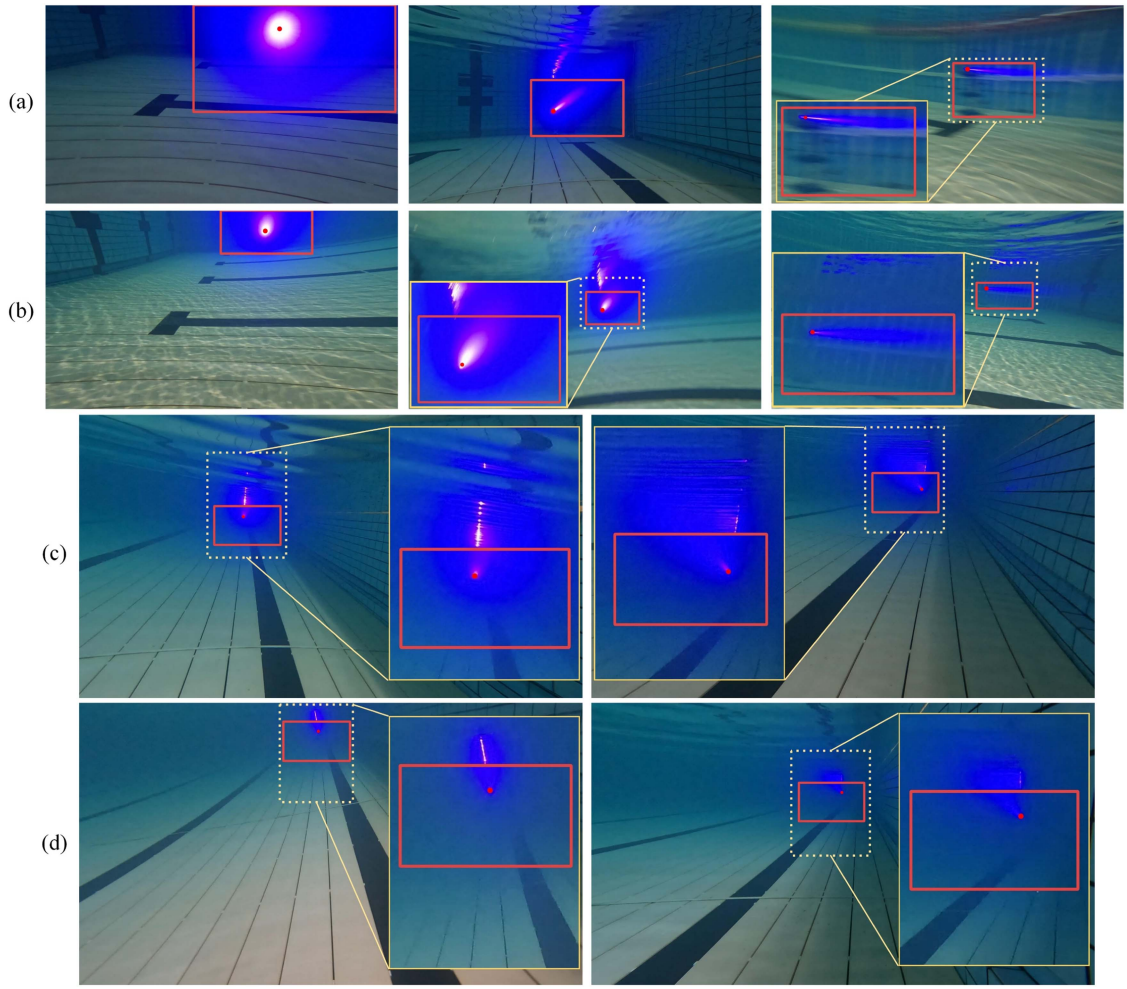

Fig. 12. Visualization results of our method. The red bounding box represents the light spot region, while the red dot represents the light source. These images were captured at distances of approximately (a) 5-m, (b) 10-m, (c) 20-m, and (d) 30-m from the light source, respectively.

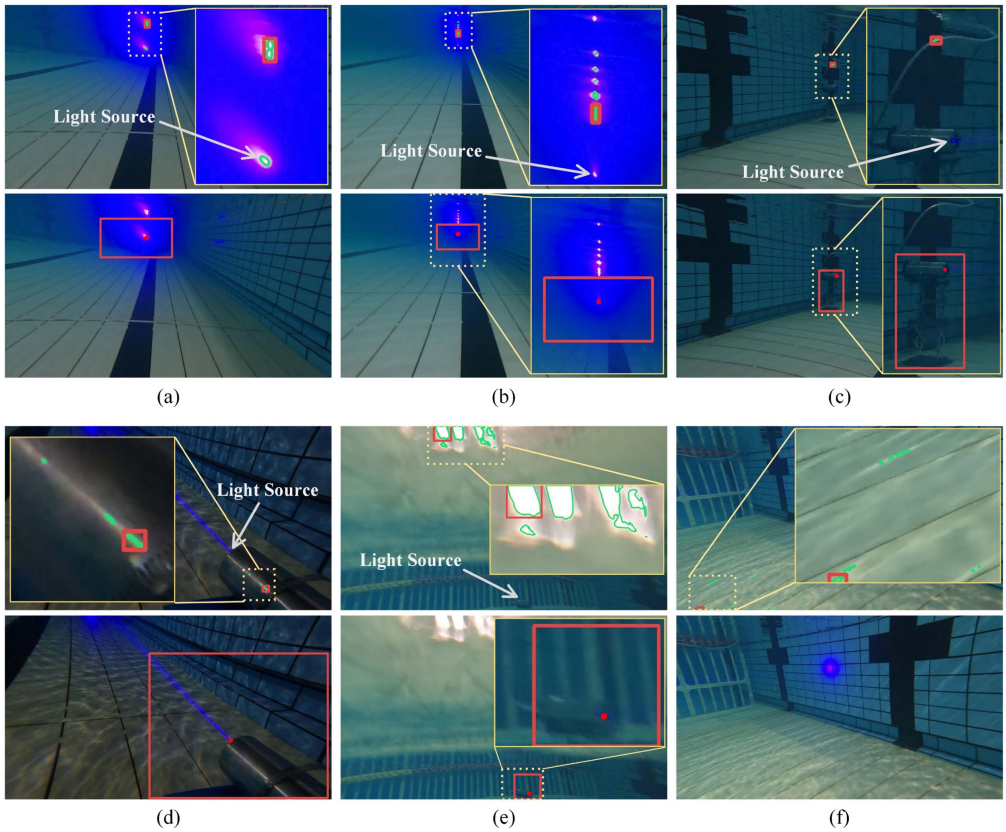

Fig. 13. Comparison with the light spot extraction method based on grayscale in scenario (a) ∼ (f). Top row: Results of grayscale-based method. The extracted light spot is marked using the green contour line. The largest spot in each image is marked by a red bounding box. Bottom row: Our results.

method. This result indicates that the high-precision light source positioning results are greatly reduced. While the RTMPose-l (single) is close to our method in several evaluation metrics, its FLOPs is up to 68.2G, which is bigger than that of our method by over eight times. In conclusion, considering both detection accuracy and computational cost, the cascaded detection method of YOLOv8s + RTMPose-t is the optimal choice.

On the ULDB test set, we compared the average positioning error of our method with the classical grayscale centroid method. In the grayscale centroid method, the light spot is first extracted from the image based on a grayscale threshold, and then the light source position is calculated based on the spot. To ensure the representativeness of the results, we extracted the light spot using different thresholds, including 180, 220, 250, and the brightest pixel value. If more than one spot is extracted from an image, only the spot with the largest area is used to calculate the light source position. As shown in Fig. 11, even in the best case, the average positioning error of the grayscale centroid method reaches 142.08 pixels. Note that many relatively dark spots are not detected in this case.

C. Qualitative Evaluation

Fig. 12 visually shows some of the results predicted by our method. It is worth noting that the images presented in this subsection are all from the test set, and the model has never "seen" them during training. It can be observed that our method works well regardless of how the terminal-camera distance changes or how the terminal deviation angle varies. In particular, when the terminal-camera distance exceeds 20 m, both the light spot region and the light source become unclear. Furthermore, there is severe reflection interference from water surface in the images. However, our method can still efficiently eliminate the interference and accurately predict the light source position.

The light source positioning algorithms presented in [17], [21], [22] all rely on the grayscale-based method for extracting light spots. The performance of this spot extraction method is visually demonstrated in Fig. 13. As described in [22], we first applied Gaussian filtering to the grayscale version of the original color image, and then the brightest area in the gray image was identified as the light spot. In Fig. 13, the light spot extracted by this method is marked by a green contour line, and the spot with the largest area in each image is marked by a red bounding box. For (a) and (b), the images are seriously disturbed by reflection, and the false light spot induced by reflection is even more dominant than the real one. Although the real spot is also extracted in (a), it is challenging to exclude interference through post-processing because the area of the real spot is smaller. Thus, the grayscale-based method cannot work in this situation. In (c)∼(e), due to the large $|\theta_d|$ of the terminal (the $|\theta_d|$ of the terminal in (d) is close to 180°), the light spot appears dim and

is no longer the brightest area of one image. As a result, the real light spot is not noticed by the grayscale-based method. In contrast, the irrelevant but brightest areas are mistakenly recognized as light spots. Scene (f) is a pure background without any terminal. In practice, such pure background images are frequently captured while the terminal rotates to search for the other terminal. However, the grayscale-based method always extracts areas with the largest gray value from an image. For a pure background image, this method fails to report the result that there is no light spot in the image and instead provides an incorrect position, which misleads the alignment system. All the above examples illustrate that the grayscale-based method is difficult to meet the practical alignment requirement.

Fortunately, our method provides accurate detection results when facing these complex scenarios. This is attributed to the powerful feature extraction capability of the DNN. During inference, the DNN not only focuses on low-level semantic information such as image intensity, but also continuously refines and combines various potential features such as shape and texture. Moreover, in principle, the deep learning method makes predictions based on whether specific features exist in the image. As a result, for a pure background image where no specific feature is present, our method can output a lower prediction score to indicate that no light spot region or light source is detected.

V. Conclusion

For the camera-assisted UAA system, the terminal-camera misalignment degree can be arbitrary before the alignment process. This implies that the position, size, brightness, and shape of light spots in captured images are all variables. Existing spot positioning methods mainly rely on simple image features and have a limited capability to handle the wide range of misalignment. In this paper, a deep learning-based cascaded light source detection method is proposed to obtain accurate light source position from various images. Thanks to the powerful feature extraction capability of the DNN, the proposed method can work well under a wide range of misalignment. The cascaded architecture provides better comprehensive performance in terms of accuracy and speed. As a necessity for deep learning, the ULDB dataset is constructed, which includes diverse misalignment scenarios. On the ULDB test set, the proposed method achieves 99.1% AP and 92.1% PCK@0.01. For the 4K images, the average positioning error of the proposed method is only 4.66 pixels, compared to 142.08 pixels for the grayscale centroid method. Additionally, our method is lightweight in the deep learning field, with 8.1G FLOPs, making it suitable for real-time operation on specialized underwater mobile devices.

In conclusion, our work presents an effective and practical solution to the alignment problem in LD-based UWOC systems. We believe that this work is merely the beginning rather than the end. There are still many interesting topics to explore in the field of light source detection. In the future, we will focus on acquiring light source image data across different water bodies with varying turbidity. Developing more accurate and faster light source detection algorithms will also be a key area of future research.

References

[1] M. F. Ali, D. N. K. Jayakody, Y. A. Chursin, S. Affes, and S. Dmitry, "Recent advances and future directions on underwater wireless communications," *Arch. Comput. Methods Eng.*, vol. 27, pp. 1379–1412, 2020.

[2] S. Zhu, X. Chen, X. Liu, G. Zhang, and P. Tian, "Recent progress in and perspectives of underwater wireless optical communication," *Prog. Quantum Electron.*, vol. 73, 2020, Art. no. 100274.

[3] C. Fei et al., "100-m/3-Gbps underwater wireless optical transmission using a wideband photomultiplier tube (PMT)," *Opt. Exp.*, vol. 30, no. 2, pp. 2326–2337, Jan. 2022.

[4] Y. Dai et al., "200-m/500-Mbps underwater wireless optical communication system utilizing a sparse nonlinear equalizer with a variable step size generalized orthogonal matching pursuit," *Opt. Exp.*, vol. 29, no. 20, pp. 32228–32243, Sep. 2021.

[5] X. Hong, C. Fei, G. Zhang, J. Du, and S. He, "Discrete multitone transmission for underwater optical wireless communication system using probabilistic constellation shaping to approach channel capacity limit," *Opt. Lett.*, vol. 44, no. 3, pp. 558–561, Feb. 2019.

[6] J. Du et al., "Experimental demonstration of 50-m/5-Gbps underwater optical wireless communication with low-complexity chaotic encryption," *Opt. Exp.*, vol. 29, no. 2, pp. 783–796, Jan. 2021.

[7] X. Hong et al., "Experimental demonstration of 55-m/2-Gbps underwater wireless optical communication using SiPM diversity reception and nonlinear decision-feedback equalizer," *IEEE Access*, vol. 10, pp. 47814–47823, 2022.

[8] Y. Guo et al., "Compact scintillating-fiber/450-nm-laser transceiver for full-duplex underwater wireless optical communication system under turbulence," *Opt. Exp.*, vol. 30, no. 1, pp. 53–69, Jan. 2022.

[9] Y. Hua et al., "Fisheye lens-based UWOC system with an FOV of±90°," *Opt. Exp.*, vol. 31, no. 16, pp. 26888–26897, Jul. 2023.

[10] J. Xiong et al., "Implementation of large field-of-view detection for UWOC systems based on a diffractive deep neural network," *IEEE Photon. J.*, vol. 15, no. 3, Jun. 2023, Art. no. 8800107.

[11] M. Zhao et al., "Long-reach underwater wireless optical communication with relaxed link alignment enabled by optical combination and arrayed sensitive receivers," *Opt. Exp.*, vol. 28, no. 23, pp. 34450–34460, Nov. 2020.

[12] M. A. Watson et al., "Assessment of laser tracking and data transfer for underwater optical communications," in *Proc. SPIE*, vol. 9248, 2014, Art. no. 92480T.

[13] Y. Weng, T. Matsuda, Y. Sekimori, J. Pajarinen, J. Peters, and T. Maki, "Establishment of line-of-sight optical links between autonomous underwater vehicles: Field experiment and performance validation," *Appl. Ocean Res.*, vol. 129, Dec. 2022, Art. no. 103385.

[14] Y. Weng, T. Matsuda, Y. Sekimori, J. Pajarinen, J. Peters, and T. Maki, "Pointing error control of underwater wireless optical communication on mobile platform," *IEEE Photon. Technol. Lett.*, vol. 34, no. 13, pp. 699–702, Jul. 2022.

[15] Y. Weng, J. Pajarinen, R. Akrour, T. Matsuda, J. Peters, and T. Maki, "Reinforcement learning based underwater wireless optical communication alignment for autonomous underwater vehicles," *IEEE J. Ocean. Eng.*, vol. 47, no. 4, pp. 1231–1245, Oct. 2022.

[16] I. Romdhane and G. Kaddoum, "A reinforcement-learning-based beam adaptation for underwater optical wireless communications," *IEEE Internet Things J.*, vol. 9, no. 20, pp. 20270–20281, Oct. 2022.

[17] A. J. Williams, L. L. Laycock, M. S. Griffith, A. G. McCarthy, and D. P. Rowe, "Acquisition and tracking for underwater optical communications," in *Proc. SPIE*, vol. 10437, 2017, Art. no. 1043707.

[18] N. D. Hardy et al., "Demonstration of vehicle-to-vehicle optical pointing, acquisition, and tracking for undersea laser communications," in *Proc. SPIE*, vol. 10910, 2019, Art. no. 109100Z.

[19] Z. Zheng, H. Yin, J. Wang, and L. Jing, "A laser spot tracking algorithm for underwater wireless optical communication based on image processing," in *Proc. IEEE 13th Int. Conf. Commun. Softw. Netw.*, 2021, pp. 192–198.

[20] J. Lin et al., "Machine-vision-based acquisition, pointing, and tracking system for underwater wireless optical communications," *Chin. Opt. Lett.*, vol. 19, no. 5, May 2021, Art. no. 050604.

[21] J. Tang, R. Jiang, Z. Chen, and Z. Zhu, "Monocular vision aided optical tracking for underwater optical wireless communications," *Opt. Exp.*, vol. 30, no. 9, pp. 14737–14747, Apr. 2022.

[22] W. Liu, Y. Jiang, N. Huang, S. Li, and Z. Xu, "Distorted laser spot center positioning based on captured image under laser-camera misalignment in UOWC," *IEEE Photon. J.*, vol. 15, no. 3, Jun. 2023, Art. no. 7302206.

[23] D. Chen et al., "Experimental study of laser spot tracking for underwater optical wireless communication," *Opt. Exp.*, vol. 32, no. 4, pp. 6409–6422, Feb. 2024.
[24] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 390–391.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[26] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023.
[27] Y. Li et al., "Simcc: A simple coordinate classification perspective for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 89–106.
[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
[29] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by ultralytics," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[30] T. Jiang et al., "RTMPose: Real-time multi-person pose estimation based on MMPose," 2023, *arXiv:2303.07399*.
[31] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.
[32] COCO dataset, "Detection evaluation," 2014. [Online]. Available: https://cocodataset.org/#detection-eval
[33] COCO dataset, "Keypoint evaluation," 2014. [Online]. Available: https://cocodataset.org/#keypoints-eval
[34] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 2637–2646.