IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 15, 2020, accepted March 8, 2020, date of publication March 11, 2020, date of current version March 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980005

Real-Time Iris Tracking Using Deep Regression Networks for Robotic Ophthalmic Surgery

HUAIYU QIU[1], ZHEN LI[2], YU YANG[2,3], CHEN XIN[4], AND GUI-BIN BIAN[2,5], (Member, IEEE)
[1]Department of Ophthalmology, Beijing Chaoyang Hospital, Capital Medical University, Beijing 100020, China
[2]State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3]School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China
[4]Beijing Tongren Eye Center, Beijing Tongren Hospital, Beijing 100730, China
[5]School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China

Corresponding author: Yu Yang (2120170273@bit.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1302704, and in part by the National Natural Science Foundation of China under Grant U1713220.

ABSTRACT Robotic-assisted platforms are expected to guarantee the accuracy of surgical operation and accelerate its learning curve. Iris tracking can guide the robotic manipulator during the operation. However, few researches focused on it during surgery. It is a big challenge due to the deformation of the iris and occlusion caused by instruments. A novel real-time iris tracking method based on a regression network are proposed to meet the speed and accuracy requirements of the ophthalmic robotic system. It utilizes the low-level visual features and high-level semantic meanings from different layers to capture the discriminative representation of the iris target. Then the bottleneck layers are added to improve computation efficiency. Furthermore, a multi-loss function is designed by jointly learning Absolute loss and Euclidean loss. Finally, the experimental results under the typical surgical scene demonstrate that iris tracker achieves an accuracy of 89.16% and a real-time speed of 134fps with GPU, which is suitable for the ophthalmic robotic system to perform real-time robotic manipulation.

INDEX TERMS Robotic surgery, deep learning, cataract surgery, iris tracking, real-time tracking.

I. INTRODUCTION

Cataracts are one of the most common ophthalmic conditions. They are the principal cause of blindness. According to a world health report, about 43% of global blindness were caused by cataracts [1]. Cataracts are largely associated with aging. Some studies have shown that 17.2%Americans older than 40 years have cataracts [2], 21.62% Chinese older than 45 have cataracts [3]. Other countries have the same trends [4].Cataracts are clouding of the eye lens. As they grow, the normal life of the patient may be affected. The specific symptoms are dim, blurred or yellow vision. Currently, surgery is the most effective and common way to treat cataracts. During cataract surgery, the surgeon will remove the cataracts and replace them with the intraocular lens (IOL) [5].

The ophthalmic robotic system is expected to automatically perform surgery in clinics. Autonomous robot can break through the human physiological limits, complete advanced procedures and increase the number of doctors [5], [6]. The typical robotic systems are precise surgical system and DA Vinci Surgical System. Retinal surgery has performed successfully in clinic assisted with The Preceyes Surgical System [7].It is the first human test of robotic eye surgery. The DA Vinci Surgical System has not been applied in clinical. It has been used for robotic-assisted pterygium surgeries in nonliving biological pterygium models [8]. However, existing ophthalmic robotic systems have not achieved autonomous operation. It only can assist with the surgeon. Other surgical fields such as urological [9], general [10], digestive [11], gynecological [12],liver [13] and cardiovascular [14] have the same problem. The main reason is that robotic vision system is not advanced enough. With the advanced vision system and tracking control approaches [15]–[17], ophthalmic robot can work in a secure and valuable manner. Building a robot with human like vision capabilities is a demanding task. It should have abilities to distinguish targets from backgrounds and identifying the moving target [18], [19]. Then it can perform the desired action through the understanding scene [20]. The ophthalmic robotic system should be able to identify the soft tissues and track them during the operation [21], [22]. Few researches focused on tissue tracking due to the challenges of the deformation, occlusion and movement of the target. To meet the speed and accuracy requirements of the robot

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

vision system, we focused on iris tracking task during the surgical operation.

Iris tracking plays a vital role in the ophthalmic robotic system. At the beginning of the cataract surgery, a very small incision was made in the cornea. The incision was applied to inject viscous material. Another very large incision was made prior to the development of phacoemulsification. The lens breaks into small pieces and the folded artificial lens are inserted into the capsule through the incision. It is ordinary to find the location of the incision for the human eye, while not easy for the robot. Manipulator should be guided by some determiner reference location. For ophthalmic robotic system, iris tracking can provide a reference location for the incision, it can guide the manipulator to move during the operation. On the other hand, the patient's eyes are difficult to remain still or not always fully visible during the surgery, which bring great challenges to the ophthalmic robotic system. Iris tracking can also play a role in protecting patients in this respect. It is a key step in intraoperative protection for the ophthalmic robotic system. Iris tracking is the indispensable part of the ophthalmic robotic system. Occlusion of surgical instruments, changes in light, and interference with drugs can also affect the implementation of robotic surgery. Iris tracking is challenging due to these inference factors. In order to perform fine cataract surgery, it is necessary to ensure that the iris location can be accurately identified. Benefit from the recording function of the binocular microscope, the surgical procedure can be saved [23]. The ophthalmic robotic system can learn from the recorded frames through deep learning methods. A large quantity of surgical frames can be annotated and trained for the iris tracker. The article aims to track the iris accurately and real-time.

The main contributions of this paper can be summarized as follows:

- A novel regression network for iris tracking during the ophthalmic surgery is proposed. To capture the discriminative representation of the iris target and take full advantage of different level features, the low-level visual features and high-level semantic meanings are fused. The Bottleneck layer and skip connection are added to the GOTURN [24] network to improve the accuracy of iris tracking.
- A multi-loss objective function is designed by jointly learning Absolute loss and Euclidean loss. The experimental results under the real surgical videos show that multi-loss network has a lower location error and the higher area overlap rate.
- The tracking speed of our method is 134fps on an NVIDIAGTX 1080Ti GPU. It demonstrates that the proposed method is suitable for the ophthalmic robotic system to perform real-time robotic manipulation.

The main body of this article is organized as follows. Section 1 is the background and the introduction of our proposed method. Section 2 elaborates on related work about iris tracking. Section 3 describes the main ideas and contains the methodologies of the experiment. The fourth section discusses the proposed method in terms of its accuracy and speed. Finally, the Section 5 concludes the paper and suggests the future work.

II. RELATED WORK

Object tracking has received much attention from a wide range of applications, such as surveillance and medical imaging over the last decades [25]. Given the interested object in a frame of a video, the goal of tracking is to locate the target in the remaining subsequent frames. Lots of trackers have achieved promising results. In this section, we discuss the representative visual trackers and iris trackers.

Early works use the handcrafted features to track the object from the images directly. Some feature descriptions are first described beginning of the tracking. Then the regions of the image are classified into different classes such as object and non-object. The object tracking is generally performed by search the candidate windows with the highest classifier score. Feature like Haar-like [26] is used for object tracking in [27]. Local Binary Patterns [28] and Scale Invariant Feature Transform (SIFT) [29] are employed to find the center of the iris [30]. The speeded-up robust features (SURF) [31] algorithm has been conducted for gaze tracking [32]. An eyelid shape model generated beforehand from PCA is used to track the iris [33]. Other researchers have shown that a template matching method is also a useful tool for iris tracking [34], [35]. These features can be efficiently extracted from images, however, have limited the ability of feature representation, which is tough to handle complex scenarios. Since the surgical scene has much occlusion of surgical instruments, handcrafted features may not be sufficient for iris tracking for the ophthalmic robotic system.

Correlation filters (CF) have become popular for visual tracking task due to its high computational efficiency. The speed results from the use of fast Fourier transforms (FFT). Frequency domain has a lower computational cost. Kernelized correlation filters (KCF) [36] computed the circulant structure and FFT for fast tracking. Various extensions of KCF have been proposed to considerably improve tracking accuracy [37], [38]. B.A. Wilson *et al.* exploited the adaptive correlation filters to track the face and eye for mobile robots [39]. The trend of CF method is to combine with CNN features [40], [41]. The biggest disadvantage of CF is a poor performance for fast deformation and fast motion, while iris deformation and motion are quite common during cataract surgery.

CNN features make a great contribution to the state-of-the-art object trackers [42]–[48] in recent years. Features learned by the neural network show strong representation of the semantic information of the target. It has attracted great interests nowadays. Bin Li presented an effective cascaded CNNs methods to detect the eye location in facial images, the first CNN can classify the region as left or right eye, the second is for detection [49]. Harini K employed ensemble learning with the ResNet10 model to track eye for iphone [50]. Wolfgang *et al.* [51] proposed a dual

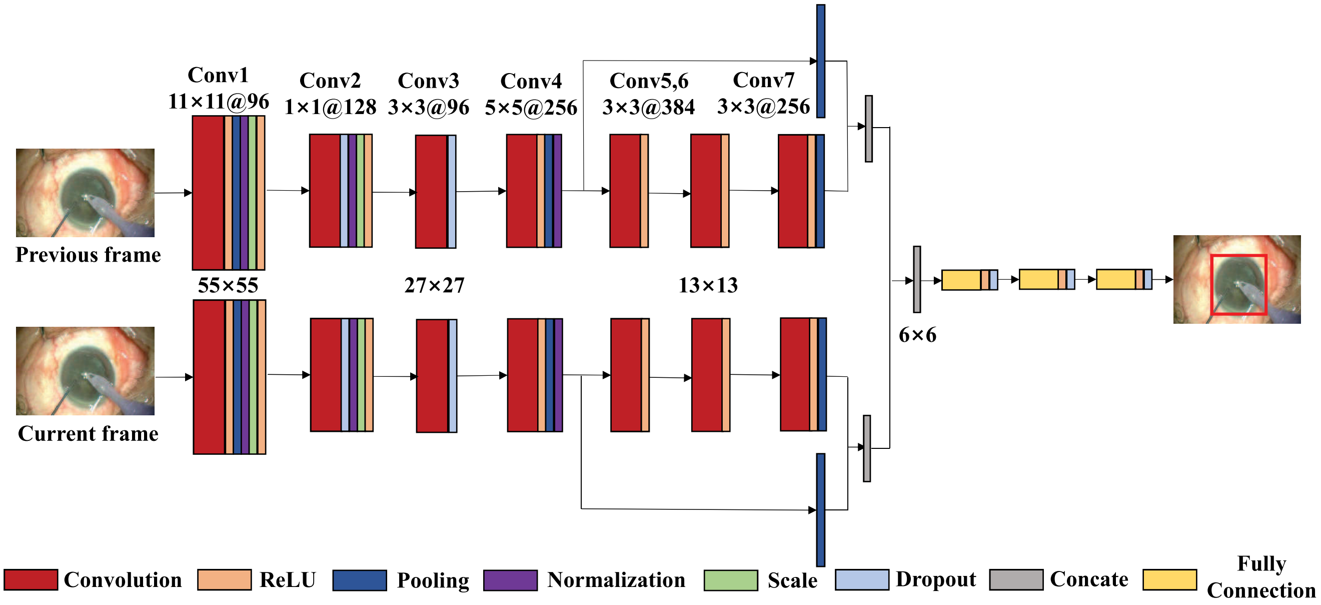


FIGURE 1. **The propose network architecture for iris tracking. It contains the feature extraction network and regression module. Different level features are integrated together.**

convolutional neural network for pupil position detection. It has two stages, the pupil position was roughly identified first and another CNN was employed to refine the position. With the development of the neural network, more and more new architectures have been proposed. 3D convolution structure is added to model the motion information of objects [52]. Semisupervised adversarial learning method is recently used in salient object detection [53], [54]. S Hoffman designed a CNN architecture for eye detection, and incorporated a segmentation mask in order to automatically learn the relative importance of the pupil and iris regions [55]. K Krafka *et al.* trained a convolutional neural network for eye tracking, while running in 10-15 fps on a modern mobile device [56]. Although deep learning methods overperform lots of hand-crafter methods and correlation filters methods on the VOT benchmark [57], most neural network-based trackers can not achieve real-time tracking due to online training.

GOTURN [24] is significantly the fastest network-based tracker, which is a state-of-the-art tracker that can run at 100fps. It uses a simple feed-forward network to train. It has learnt a generic relationship between object motion and appearance. It can be employed to track novel objects. Although the GOTURN [24] method has a very satisfactory throughput, it only utilizes the high-level features and is trained with generic objects. The Bottleneck layer and skip connection are added to the GOTURN [24] network to improve the accuracy of iris tracking. Multi-loss objective function is designed to get a better overlap precision. These modifications significantly improve the track accuracy for the ophthalmic robotic system.

III. METHOD

The proposed method will be explained in detail in this section. The architecture of the proposed network is illustrated in Fig.1. It begins by utilizing a feature extraction network to produce a set of feature maps, then different level features are integrated. The low-level visual features and high-level semantic meanings from different layers can be utilized to capture the discriminative representation of the iris target. After that the loss objective function is computed between the ground truth and the predicted bounding box. Here a new multi-loss function is designed by joint learning Absolute loss and Euclidean loss. Finally, the target location is regressing. In following subsections, we first introduce the network architecture, including the feature integration method and the designed multi-loss function, and then elaborate the training process.

A. NETWORK ARCHITECTURE

1) INPUT

Our network takes the two frames as input. It contains the current frame image and the next frame image. It can output the localization result in an end-to-end manner. The size of the input image is 256 × 256. The target localization is marked on the first frame. We use four parameters $R_t = \{x, y, w, h\}$ to describe the target localization. Where R_t is the target region, x, y stand for the center coordinate of the iris target, w is the width of the target bounding box, and h is the height of the bounding box. Due to the smooth movement of the target, previous localization can provide a reference for the current frame [24]. The current frame is cropped according the previous target localization. $R_s = \{x_s, y_s, w_s, h_s\}$. We set:

$$\begin{aligned} x_s &= x, \quad y_s = y \\ w_s &= 2 \times w \\ h_s &= 2 \times h \end{aligned} \tag{1}$$

where R_s is the search region, x_s, y_s stand for the center coordinate of the search region, w_s is the width of the search region, and h_s is the height of the bounding box. Then Two region are fed to the CNN, respectively.

2) FEATURE EXTRACTION AND REGRESSION

The network is composed of convolution, pooling, concatenation and Fully-Connected (FC) layers. A set of CNN features are generated from two cropped inputs. Each CNN has 7 convolution layers. The convolution layer produces the linear feature map:

$$f_s(t) = k_s * t + b_s \tag{2}$$

where k_s and b_s stand for the convolution kernel weight and bias, $*$ is the convolution operator. We apply multi-scale convolution operation to capture multi-scale contextual information. It contains $1 \times 1, 3 \times 3, 5 \times 5, 11 \times 11$ convolution layers. Each CNN extracts different hierarchical features from its ROIs. The 1×1 convolution layer combines the feature maps for information integration. Other size convolution layers reduce the resolution of the feature map. Then the different level features are integrated together to represent the iris target. The low-level visual features and high-level semantic meanings from different layers can be both utilized to capture the discriminative representation of the target. The relu layer is used to generate non-linear feature map:

$$f_k(t) = ReLU(bn(f_s(t))) \tag{3}$$

where *bn* stands for the batch normalization. Furthermore, the features of two branches are concatenated and fed to 3 FC layers. Finally, the fully connected layer regresses the localization of the iris target. The output of the linear regression layer is:

$$f(x_i) = w_0 h(x_i) + b_0 \tag{4}$$

where $h(x_i)$ is the output of the previous concatenated pooling layer. The output of the network has four components corresponding to the x and y coordinates of the upper corner, the width and height of the bounding box.

3) BOTTLENECK LAYERS

The most significant thing is to ensure real-time tracking for ophthalmic robotic system, Bottleneck layers can improve computational efficiency. GOTURN [24] only used the convolutional layers of AlexNet [58] as the feature extraction network. Referring to the DenseNet [59],ResNet [60] and Highway NetWorks [61], that a 1×1 convolution layer can be introduced as bottleneck layer before 3×3 convolution to improve computational efficiency. It also can combine the feature maps for information integration. In our network, we add 1×1 convolution layer after the first pooling layer.

4) SKIP CONNECTION

Skip Connections help CNN directly extract coarse high-level semantic feature and low-level visual features from different layers. Multilayer features can be fused to capture both simple features and semantic features. As opposed to the original GOTURN [24] network that only use the high-level features, low-level features and high-level features are combined in the proposed network. It is shown in Fig.2. Different level layers extract different features. The shallow layers of CNN extract simple visual and high-resolution features, such as corners and edges. The deep layers contain semantic information. Both low-level and high-level features make contributions to capture the discriminative representation of the iris target. We use skip connection to ensemble multilayer features together to train our network. Different level features are concatenated to improve the tracking accuracy. Multi-level features help the CNN generate multi-scale features with accurate spatial and semantic information closely to iris tracking task.

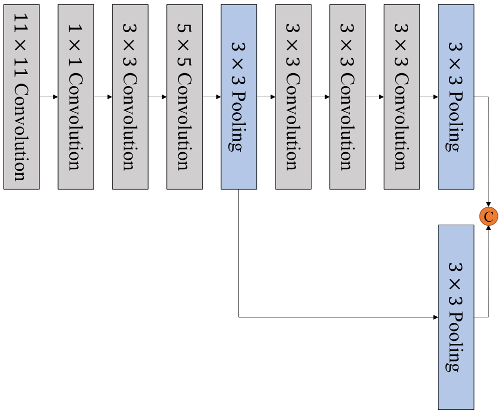


FIGURE 2. The feature fusion architecture.

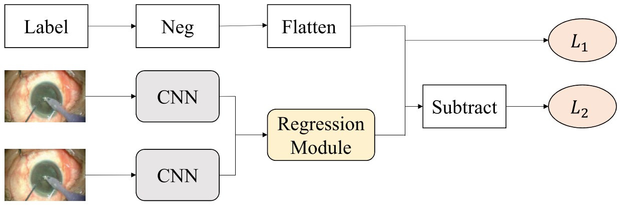


FIGURE 3. The proposed multi-loss function framework.

5) MULTI-LOSS

A multi-loss function is designed to reduce the iris location error. It can jointly learn the Absolute loss and Euclidean loss with different weight. It is shown in Fig.3. The iris position can be accurately localized with the multi-loss function. It is defined as:

$$L = \lambda_1 L_1 + \lambda_2 L_2 \tag{5}$$

where L_1 is Absolute loss and L_2 is Euclidean loss. Here λ is a constant coefficient balancing the two parts of the two loss. In our experiment,$\lambda_1 = 0.7$, $\lambda_2 = 0.3$.

$$L_1 = \sum_{n=1}^{N} |\hat{y_n} - y_n| \tag{6}$$

$$L_2 = \frac{1}{2N} \sum_{n=1}^{N} \|\hat{y_n} - y_n\|_2^2 \tag{7}$$

where $\hat{y_n}$ represents the label and y_n represents the network output of the given input data. N is the batch size.

B. TRAINING

Training of the proposed network is the same as GOTURN [24] model. The proposed tracker is initialized with the bounding box in the first frame. Transfer learning is used and iris tracking model is fine-tuned from pre-trained imagenet model. The pre-trained model has learned a discriminative feature representation of the generic object, which can help to solve the iris tracking task. We train our network with Caffe [62]. The 1×1 convolution layer is initialized by msra distribution and other convolution layers initialized by Gaussian distribution. The base learning rate is 1e-6, the learning policy is step.

A dataset of 13 videos from 13 consecutive cataract surgeries was collected at Beijing Chaoyang Hospital and Beijing Tongren Hospital (Beijing, China) between April and June 2018. A total of 17221 frames of 7 patients were labeled as training dataset. The training data includes the continuous procedures in cataract surgery. In our experiment, per subclasses in the training set are all according to the procedures in the real surgery. Each patient's video contains the main 6 steps of the cataract surgery, including liquid injection, incision, capsulorhexis, phacoemulsification, aspiration of lens and intraocular lens implant process. In addition, external interference is also considered in the data set, videos that contain deformation, movements and occlusion are also in the training set. The remaining 6 videos are used to evaluate the performance of our tracker. There are 4531 frames for test. It contains the main 6 steps and 3 external interference scenes.

We have written a labeling program using Matlab software to label the training set and test set. The targets in each frame of video are labeled with rectangular boxes, and the position coordinates of the boxes are saved in the txt file.In the testing stage, tracker predict the bounding box of the target for every frame.

Algorithm 1 summarizes the detail procedure of our proposed iris tracking algorithm. It contains four parts: crop, according to the initial frame, training using the multi-loss function, optimization and test. Since iris usually move smoothly, our tracking algorithm first crop the two input frame. It makes the tracking algorithm less computation. Secondly, two inputs are forward propagated through the proposed network. Both low-level and high-level features make contributions to capture the discriminative representation of the iris target. Feature integration can improve the tracking accuracy. Then we design a multi-loss functions to optimize the network parameters. It can jointly learn Absolute loss and Euclidean loss. It can produce lower location error and higher area overlap rate. After training procedure, the weights are frozen, test, videos are forward propagated through the trained network. Finally, the network outputs the iris localization.

Algorithm 1

```
Training data $D$, parameters $\theta$, hyperparameters $\hat{\theta}$
**Input: Previous frame and current frame**
**Output: Iris localization**
**STEP 1: Crop according to the initial frame**
Generate the search region $R_s = \{x_s, y_s, w_s, h_s\}$
    according to the initial target region.
$D_c \Leftarrow crop(D)$
**STEP 2: Training using multi-loss function**
Two frame is forward propagated through the proposed
    feature extraction and regression network.
**STEP 3: Optimization**
Optimizing the parameters using the multi-loss $L$ as
    equation 5.
**for** $epoch \in [0, max\_epochs]$ **do**
    **for** $i \in [0, num\_batches]$ **do**
        $L_{batch} = 0$
        **for** $d_c \in D_c$ **do**
            $L \Leftarrow \lambda_1 L_1(d_c; \theta) + \lambda_2 L_2(d_c; \theta)$
            $L_{batch} = L_{batch} + L$
        **end for**
        $\theta \Leftarrow Backpro(L_{batch}, \theta, \hat{\theta})$
    **end for**
    $\hat{\theta} \Leftarrow ParameterUpdate(epoch, \hat{\theta})$
**end for**
**STEP 4: Test**
Input a test cataract video, forward propagate the data
    through the network with trained weights, record the
    output target localization.
```

IV. EXPERIMENT AND ANALYSIS

In this section, the effect of the proposed method will be explained through experimental results. Iris tracker is implemented on python 2.7 and Caffe [62]. It runs at 134fps in a single NVIDIA GTX 1080Ti GPU with CuDNN v5.1. In our system, the base learning rate is 0.000001, the learning policy is step, the step size is 100000, the momentum is 0.9. In the comparison of multiple training processes, the above parameters can ensure training convergence.

Experiments are conducted to verify the effectiveness of the proposed iris tracking method.

Firstly, we use the test dataset to evaluate the tracking performance of the proposed method. The tracking result is compared to the marked bounding box. Not only the tracking results are shown, but also the center point coordinate of the iris target is compared.

Secondly, the advantages of the feature fusion method and bottleneck layers are analyzed. We compare the network that use single loss function with GOTURN [24] network. Two methods use the same loss function in this aspect. Then, the availability of the designed multi-loss function is demonstrated. We compare the multi-loss method with the single loss method. Two methods use the same convolution layers in this aspect.

Finally, speed comparisons are conducted in the same hardware environment.

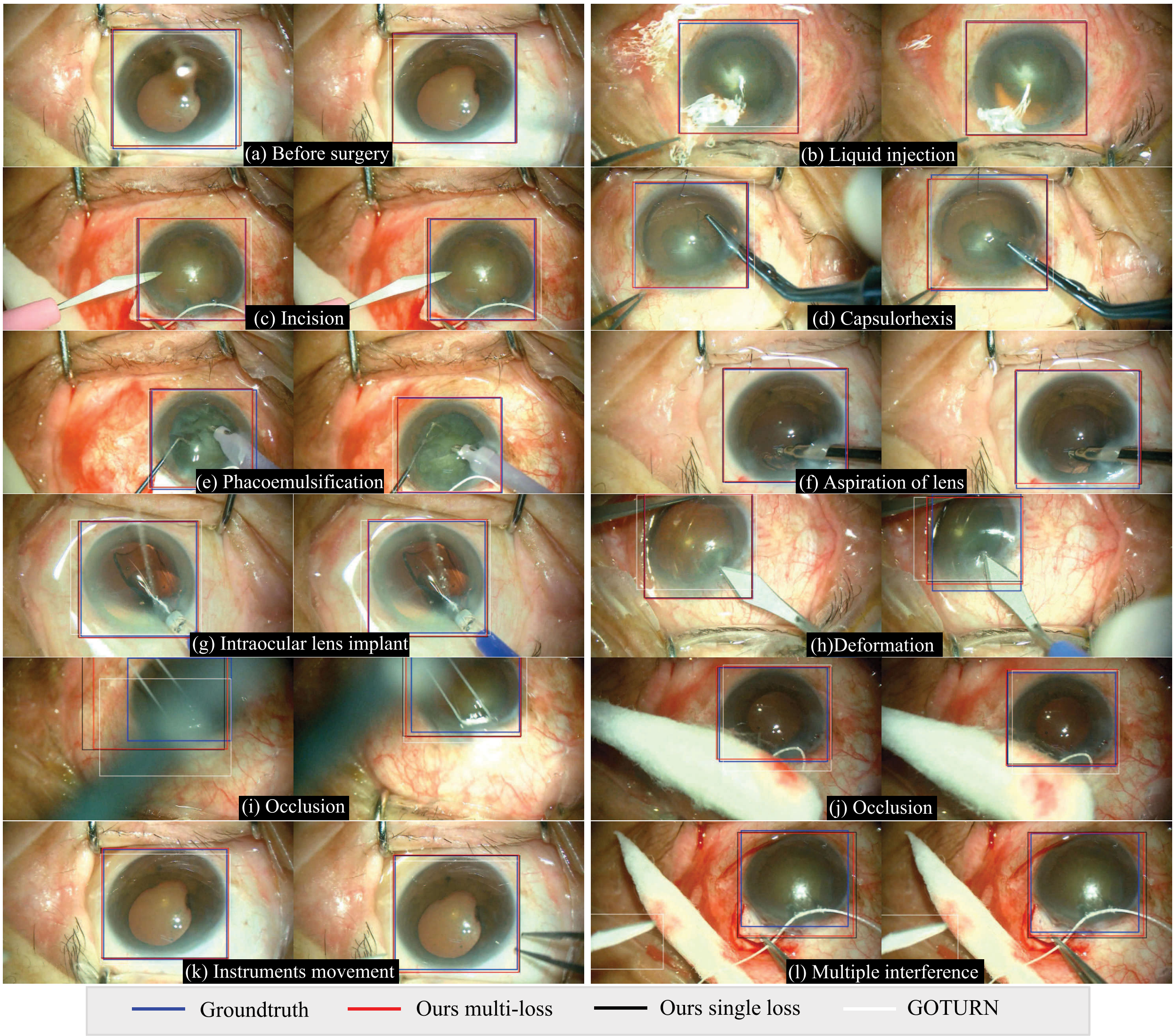


FIGURE 4. Some tracking results of the proposed tracker. The blue box stands for the ground truth. The red box stands for the tracking results of our multi-loss method.The black box stands for the tracking results of our single loss method. The white box stands for the tracking results of GOTURN method.

A. TRACKING PERFORMANCE

Some tracking results are shown in fig. 4. We have used two metrics: precision plot and success plot to validate the accuracy of the proposed method. Both two metrics are common evaluation indicators in the field of target tracking. Experimental results are also presented in two forms: tracking video with tracking box, accuracy and speed comparison data. We choose the typical surgery scenes to validate the proposed method. It contains the main procedure of the cataract surgery. The first to sixth row represent the main 6 steps of the cataract surgery, including liquid injection, incision, capsulorhexis, phacoemulsification, aspiration of lens and intraocular lens implant process. The last three rows stand for the deformation, movements and occlusion process that may occur during surgery.

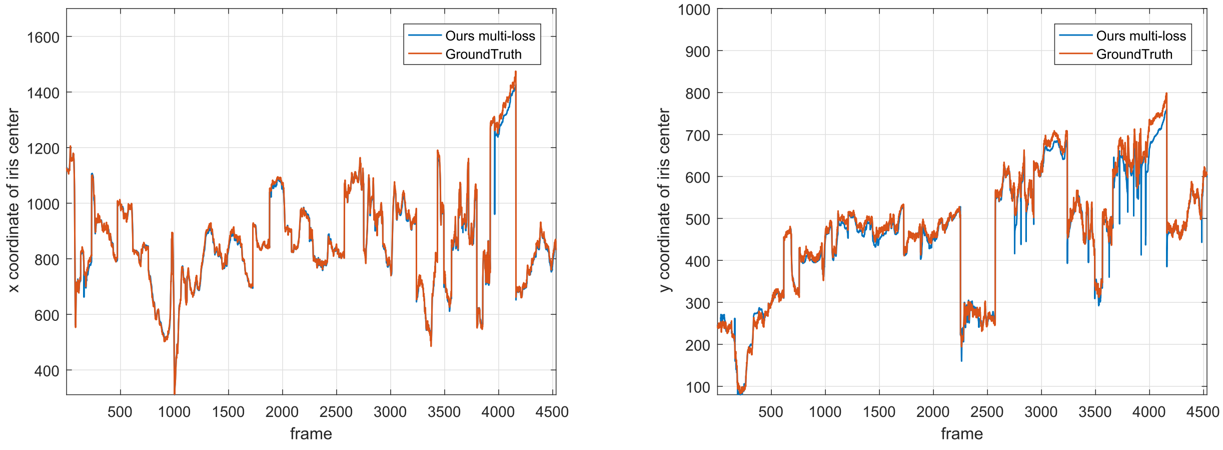
Ophthalmologists may use different color and different type instruments to perform surgery. The injection of the liquid may cause reflection of light, which will bring challenges to the robotic vision. The surgical operation will also cause the deformation and occlusion of the iris target. The iris tracker must robust enough to face various conditions for the ophthalmic robotic system. We compare the tracking results with the labeled bounding box in Fig.4. We can observe that our method can track the iris accurately under the different challenge surgical scenes.

We also compare the coordinates of the target center point with the label in fig. 5. There are totally 4531 test frames. It contains the various typical surgery scenes shown in Fig.4. It is observed that the proposed method can accurately track the iris in both horizontal and vertical directions. Our tracking results are quite close to the ground truth.

B. ACCURACY COMPARISON

We use two metrics: precision plot and success plot [57] to validate the accuracy of the proposed method.

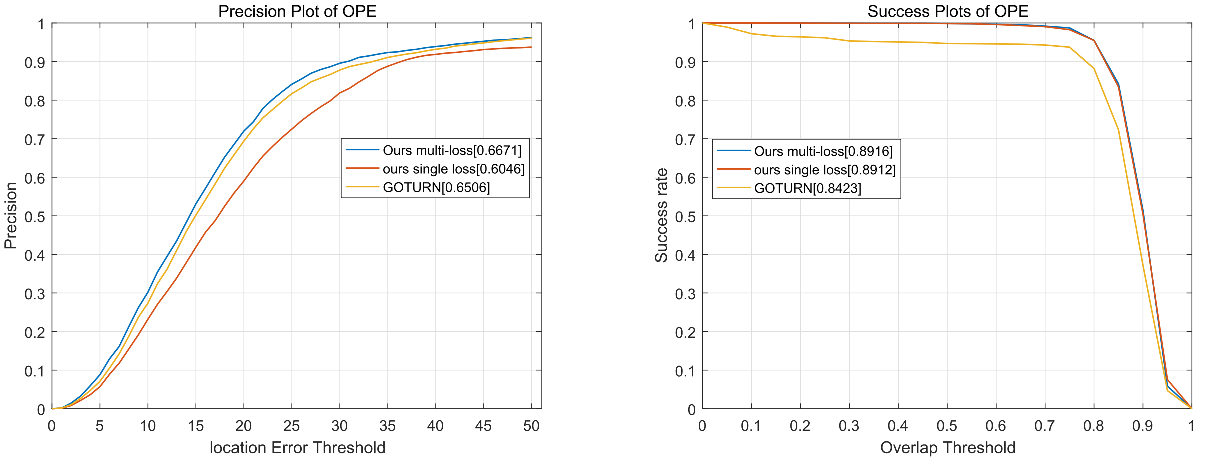
Firstly, in order to demonstrate the advantages of the feature fusion method and bottleneck layers, we compare the network with GOTURN [24] network. Two methods use the



(a) x coordinate of iris center

(b) y coordinate of iris center

FIGURE 5. Iris center comparison with groundtruth.



(a) Precision plot

(b) Success plot

FIGURE 6. Overall comparison.

same loss function. We use "Ours single loss" to stand for the method that employ the proposed convolution architecture with the Absolute loss function. We use "Ours multi-loss" to stand for our method that use the proposed convolution architecture with the multi-loss function.

Secondly, to prove the effectiveness of the designed motels function, we compare the multi-loss method with the single loss method. Two methods use the same convolution layers.

We have selected 9 typical scenes that have the most representative role in the surgical process to analyze the accuracy of the tracking algorithm. The overall comparison for three trackers is shown in Fig.6. The 9 typical scenes results comparison for three trackers are shown in Fig.7. The values in the legend are the Area under curve (AUC) scores for precision and success plots. The higher AUC value indicates better performance. According to the comparative data, the algorithm proposed in this paper performs best in 6 of these scenes, especially in liquid injection, deformation, and instrument movement. That is, in the presence of external interference, the algorithm can still achieve accurate tracking. From a clinical perspective, the target tracking algorithm should have the ability to adapt to external interference, so the method in this paper can assist the ophthalmic robotic system to achieve precise operations.

TABLE 1. The accuracy and speed comparison.

Method	Auc	Prec	Speed(fps)
Ours multi-loss	**0.8916**	**0.6671**	134.67
Ours single loss	0.8912	0.6046	136.48
GOTURN [25]	0.8423	0.6506	**186.01**

We also list the accuracy values in Table 1 and detailed values of typical scenes in Table 2. Due to the low-level visual features and high-level semantic feature are integrated, single loss method has higher accuracy than GOTURN [24] method. Due to the designed multi-loss function, multi-loss method has higher overlap and lower localization error than the single loss method. Among the above methods, our method achieves the best performance.

C. SPEED COMPARISON

The results of speed comparison are shown in Table 1. Since the new layers and extra loss function are added, the speed of our method is slightly slower than the original

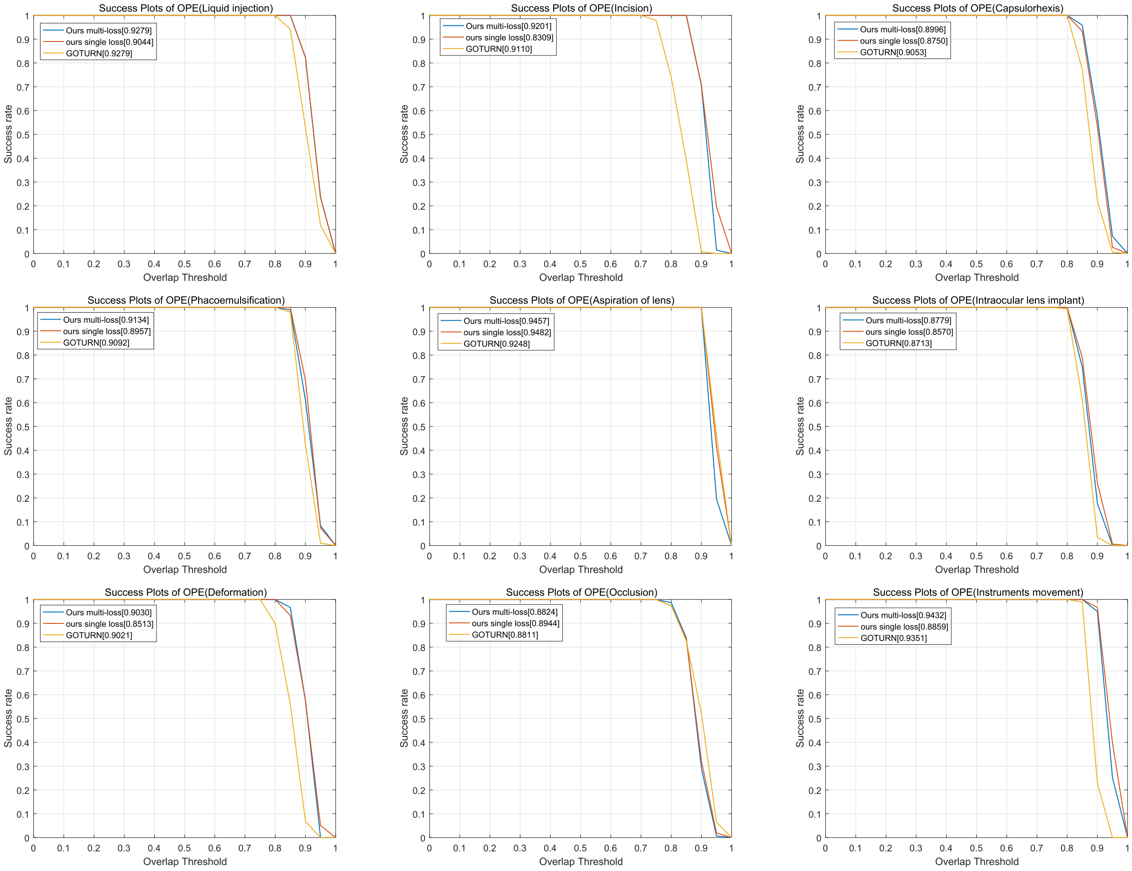


FIGURE 7. Success plot comparison of 9 typical test scenes.

TABLE 2. The accuracy comparison of 9 typical scenes.

Scenes	Ours multi-loss	Ours single loss	GOTURN
Liquid injection	**0.9201**	0.8309	0.9110
Incision	**0.9314**	0.8957	0.9092
Capsulorhexis	0.8824	**0.8944**	0.8811
Phacoemulsification	**0.9432**	0.8859	0.9351
Aspiration of lens	**0.9030**	0.8513	0.9021
Intraocular lens implant	0.8996	0.8750	**0.9053**
Deformation	**0.9279**	0.9044	**0.9279**
Occlusion	0.9457	**0.9482**	0.9348
Instruments movement	**0.8779**	0.8570	0.8713

GOTURN network. But it can also achieve real-time tracking. The tracking speed can meet the requirements of the ophthalmic robotic system.

GOTURN [24] is the fastest network-based tracker that can run at 100fps with ordinary GPU. It just includes a simple feed-forward network. No extra online training is needed. It has learnt a generic relationship between object motion and appearance through the training procedure. The trained model can be employed to track novel objects that not included in the training dataset. But it has a lower tracking accuracy since it only uses the high-level feature.

The Bottleneck layer and skip connection are added to the GOTURN [24] network to improve the accuracy of iris tracking. The low-level visual features and high-level semantic meanings are both utilized from different layers in our method. It can capture the discriminative representation of the iris target. From Table 1, compare with the GOTURN [24] method, we can observe that the model using fused features has higher accuracy. Compare with the same network, which only use a single loss function, the results show that the model with the designed multi-loss function has a higher AUC score for precision and success plot. The experimental results under the typical surgical scene demonstrate that the proposed method can track the iris accurately in real time.

Iris tracking can help ophthalmic robotic system work in a secure and valuable manner. The ophthalmic robotic system should be able to identify the soft tissues and track them during the operation. Iris tracking can provide a reference location for it and guide its manipulator to move during the operation. It can also play a role in protecting patients.

It is necessary to ensure that the iris location can be accurately identified for the ophthalmic robotic system. However, there are still some problems, such as when the liquid is injected or there is serious occlusion, the tracking accuracy is not high enough, which is the next step to improve.

V. CONCLUSION

The aim of this study is to develop an iris tracker specifically for the ophthalmic robotic system. For this purpose, a real-time CNN tracker is proposed. Iris tracking can provide a reference location for the incision and also play a key role in protecting patients.

In the proposed structure, different hierarchical features are integrated. The tracker can learn more discriminating features by adding the bottleneck layer and skip connection. Furthermore, to accurately localize the iris position, not only the different level features are fused, but also a multi-loss function is designed by jointly learning Absolute loss and Euclidean loss. The experimental results under the typical surgical scene demonstrate that iris tracking accuracy achieves 89.16% at 134fps. Our method outperforms the original GOTURN network. The proposed scheme is general enough to be adapted to the real surgical scene. It can track the iris accurately and real-time, it is suitable for the ophthalmic robotic system to perform real-time robotic manipulation.

Deep learning-based studies are still going on for robotic assisted surgery, further investigations include joint iris tracking and key point detection task, and then convey control commands to the robot according the detection results.

REFERENCES

[1] *World Health Report: Life in the 21st Century—A Vision for All*, World Health Org., Geneva Switzerland, 1998, pp. 391–392, no. 3.

[2] N. Congdon, J. R. Vingerling, B. E. Klein, S. West, D. S. Friedman, J. Kempen, B. O'Colmain, S. Y. Wu, and H. R. Taylor, "Prevalence of cataract and pseudophakia/aphakia among adults in the united states," *Arch. Ophthalmol*, vol. 122, no. 4, pp. 487–494, 2004.

[3] P. Song, H. Wang, E. Theodoratou, K. Y. Chan, and I. Rudan, "The national and subnational prevalence of cataract and cataract blindness in China: A systematic review and meta-analysis," *J. Global Health*, vol. 8, no. 1, Jun. 2018, Art. no. 010804.

[4] P. Mitchell, R. G. Cumming, K. Attebo, and J. Panchapakesan, "Prevalence of cataract in Australia: The blue mountains eye study," *Ophthalmology*, vol. 104, no. 4, pp. 581–588, 1997.

[5] W. Liu, Y. Su, W. Wu, C. Xin, Z.-G. Hou, and G.-B. Bian, "An operating smooth man–machine collaboration method for cataract capsulorhexis using virtual fixture," *Future Gener. Comput. Syst.*, vol. 98, pp. 522–529, Sep. 2019.

[6] A. K. Sangaiah, D. V. Medhane, G.-B. Bian, A. Ghoneim, M. Alrashoud, and M. S. Hossain, "Energy-aware green adversary model for cyberphysical security in industrial system," *IEEE Trans Ind. Informat.*, vol. 16, no. 5, pp. 3322–3329, May 2020.

[7] R. MacLaren, "First human test of robotic eye surgery a success," Tech. Rep., 2018.

[8] T. Bourcier, M. Nardin, A. Sauer, D. Gaucher, C. Speeg, D. Mutter, J. Marescaux, and P. Liverneaux, "Robot-assisted pterygium surgery: Feasibility study in a nonliving porcine model," *Transl. Vis. Sci. Technol.*, vol. 4, no. 1, p. 9, Jan. 2015.

[9] D. Murphy, M. S. Khan, P. Dasgupta, and B. Challacombe, "Robotic technology in urology," *Postgraduate Med. J.*, vol. 82, no. 973, pp. 743–747, 2006.

[10] I. Gkegkes, I. Mamais, and C. Iavazzo, "Robotics in general surgery: A systematic cost assessment," *J. Minimal Access Surgery*, vol. 13, no. 4, p. 243, 2017.

[11] P. Alberto, A. Pietro, and B. Nicolas, "Advanced applications of robotics in digestive surgery," *Transl. Med.*, vol. 1, pp. 21–50, Sep./Dec. 2011.

[12] J. Knight and P. F. Escobar, "Cost and robotic surgery in gynecology," *J. Obstetrics Gynaecol. Res.*, vol. 40, no. 1, pp. 12–17, Jan. 2014.

[13] Q. Huang, G.-B. Bian, X.-G. Duan, H.-H. Zhao, and P. Liang, "An ultrasound-directed robotic system for microwave ablation of liver cancer," *Robotica*, vol. 28, no. 2, pp. 209–214, Mar. 2010.

[14] M. Pettinari, E. Navarra, P. Noirhomme, and H. Gutermann, "The state of robotic cardiac surgery in europe," *Ann. Cardiothoracic Surgery*, vol. 6, no. 1, pp. 1–8, Jan. 2017.

[15] P. Liu, H. Yu, and S. Cang, "Adaptive neural network tracking control for underactuated systems with matched and mismatched disturbances," *Nonlinear Dyn.*, vol. 98, no. 2, pp. 1447–1464, Oct. 2019.

[16] P. Liu, M. N. Huda, Z. Tang, and L. Sun, "A self-propelled robotic system with a visco-elastic joint: Dynamics and motion analysis," *Eng. Comput.*, pp. 1–15, Feb. 2019.

[17] P. Liu, H. Yu, and S. Cang, "Trajectory synthesis and optimization of an underactuated microrobotic system with dynamic constraints and couplings," *Int. J. Control, Autom. Syst.*, vol. 16, no. 5, pp. 2373–2383, Oct. 2018.

[18] D. Zang, G.-B. Bian, Y. Wang, and Z. Li, "An extremely fast and precise convolutional neural network for recognition and localization of cataract surgical tools," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, 2019, pp. 55–64.

[19] R.-J. Huang, G.-B. Bian, C. Xin, Z. Li, and Z.-G. Hou, "Path planning for surgery robot with bidirectional continuous tree search and neural network," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 3302–3307.

[20] T. Xu, J. Yu, C.-I. Vong, B. Wang, X. Wu, and L. Zhang, "Dynamic morphology and swimming properties of rotating miniature swimmers with soft tails," *IEEE/ASME Trans. Mechatronics*, vol. 24, no. 3, pp. 924–934, Jun. 2019.

[21] X. Wu, J. Liu, C. Huang, M. Su, and T. Xu, "3-D path following of helical microswimmers with an adaptive orientation compensation model," *IEEE Trans. Autom. Sci. Eng.*, to be published.

[22] T. Xu, Y. Guan, J. Liu, and X. Wu, "Image-based visual servoing of helical microswimmers for planar path following," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 1, pp. 325–333, Jan. 2020.

[23] B. Raju, N. D. Raju, A. Raju, C. Sudhakaran, and A. Razak, "Digital video recording and archiving in ophthalmic surgery," *Indian J. Ophthalmol.*, vol. 54, no. 1, pp. 53–57, 2006.

[24] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 fps with deep regression networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 749–765.

[25] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.

[26] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, p. 1.

[27] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 983–990, 2009.

[28] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[29] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[30] W. Zhang, M. L. Smith, L. N. Smith, and A. Farooq, "Gender and gaze gesture recognition for human-computer interaction," *Comput. Vis. Image Understand.*, vol. 149, pp. 32–50, Aug. 2016.

[31] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 404–417, 2006.

[32] H. Heo, W. O. Lee, J. W. Lee, K. R. Park, E. C. Lee, and M. Whang, "Object recognition and selection method by gaze tracking and SURF algorithm," in *Proc. Int. Conf. Multimedia Signal Process.*, vol. 1, May 2011, pp. 261–265.

[33] K. Tamura, K. Hashimoto, and Y. Aoki, "Head pose-invariant eyelid and iris tracking method," *Electron. Commun. Jpn.*, vol. 99, no. 2, pp. 19–27, 2016.

[34] A. Bukhalov and V. Chafonova, "An eye tracking algorithm based on Hough transform," in *Proc. Int. Symp. Consum. Technol. (ISCT)*, May 2018, pp. 49–50.

[35] F. Wang, X. Chen, D. Wang, and B. Yang, "An improved image-based iris-tracking for driver fatigue detection system," in *Proc. 36th Chin. Control Conf. (CCC)*, Jul. 2017, pp. 11521–11526.

[36] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, *Exploiting the Circulant Structure of Tracking-by-Detection with Kernels*. Berlin, Germany: Springer, 2012.

[37] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[38] T. Liu, G. Wang, and Q. Yang, "Real-time part-based visual tracking via adaptive correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4902–4912.

[39] V. D. My and A. Zell, "Real time face tracking and pose estimation using an adaptive correlation filter for human-robot interaction," in *Proc. Eur. Conf. Mobile Robots*, Sep. 2013, pp. 119–124.

[40] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3074–3082.

[41] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5000–5008.

[42] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zuo, C. Shen, R. W. H. Lau, and M.-H. Yang, "VITAL: VIsual tracking via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8990–8999.

[43] D. Martin, R. Andreas, K. F. Shahbaz, and F. Michael, *Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking*. Springer, 2016.

[44] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg, "Unveiling the power of deep tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 483–498.

[45] T. Carneiro, R. V. Medeiros Da Nobrega, T. Nepomuceno, G.-B. Bian, V. H. C. De Albuquerque, and P. P. R. Filho, "Performance analysis of Google colaboratory as a tool for accelerating deep learning applications," *IEEE Access*, vol. 6, pp. 61677–61685, 2018.

[46] Z. Gao, H. Zhang, S. Dong, S. Sun, X. Wang, G. Yang, W. Wu, S. Li, and V. H. C. D. Albuquerque, "Salient object detection in the distributed cloud-edge intelligent network," *IEEE Netw.*, to be published.

[47] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[48] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[49] B. Li and H. Fu, "Real time eye detector with cascaded convolutional neural networks," *Appl. Comput. Intell. Soft Comput.*, vol. 2018, pp. 1–8, Apr. 2018.

[50] H. Kannan, "Eye tracking for the iPhone using deep learning," Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep., 2017.

[51] F. Wolfgang, S. Thiago, K. Gjergji, and K. Enkelejda, "PupilNet: Convolutional neural networks for robust pupil detection," *Revista De Odontologia Da Unesp*, vol. 19, no. 1, pp. 806–821, 2016.

[52] S. Dong, Z. Gao, S. Pirbhulal, G.-B. Bian, H. Zhang, W. Wu, and S. Li, "IoT-based 3D convolution for video salient object detection," *Neural Comput. Appl.*, vol. 32, no. 3, pp. 735–746, Feb. 2020.

[53] C. Wang, S. Dong, X. Zhao, G. Papanastasiou, H. Zhang, and G. Yang, "SaliencyGAN: Deep learning semisupervised salient object detection in the fog of IoT," *IEEE Trans Ind. Informat.*, vol. 16, no. 4, pp. 2667–2676, Apr. 2020.

[54] H. Zhang, Z. Gao, L. Xu, X. Yu, K. C. L. Wong, H. Liu, L. Zhuang, and P. Shi, "A meshfree representation for cardiac medical image computing," *IEEE J. Transl. Eng. Health Med.*, vol. 6, 2018, Art. no. 1800212.

[55] S. Hoffman, R. Sharma, and A. Ross, "Convolutional neural networks for iris presentation attack detection: Toward cross-dataset and cross-sensor generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1620–1628.

[56] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2176–2184.

[57] S. J. Hadfield, K. Lebeda, and R. Bowden, "The visual object tracking VOT2014 challenge results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Sep. 2014, pp. 1949–1972.

[58] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[59] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[61] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[62] J. Yangqing, S. Evan, D. Jeff, K. Sergey, L. Jonathan, G. Ross, G. Sergio, and D. Trevor, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

HUAIYU QIU was born in November 1975. He received the Ph.D. degree from the Beijing Chaoyang Hospital, Capital Medical University. He is currently the Vice Director and an Ophthalmologist with the Beijing Chaoyang Hospital, Capital Medical University. He has been engaged in ophthalmology for more than 20 years with the Lianyungang Donghai People's Hospital, the Beijing Military General Hospital, and the PLA General Hospital. He is also a member of the Neuro-Ophthalmology Group, Ophthalmology Branch, Chinese Medical Association; the neuro-Ophthalmic Branch, Ophthalmic Committee, Chinese Medical Doctor Association; and the Ophthalmology Branch, Youth Committee, Beijing Medical Association.

ZHEN LI received the bachelor's and master's degrees in mechanical engineering and automation from Beihang University, Beijing, China, in 2011 and 2014, respectively. She is currently an Associate Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing. Her research interest includes design and control of surgery robots.

YU YANG received the bachelor's degree in telecommunication engineering from Jilin University, in 2017. She is currently pursuing the master's degree with the Beijing Institute of Technology. Her current research interests include image processing and deep learning.

CHEN XIN graduated from the Peking University Health Science Center, in 2004. She received the Ph.D. degree from Capital Medical University, in 2016. She has worked as a General Ophthalmologist for five years and specialized in glaucoma and cataract for six years. She is currently the Associate Chief Ophthalmologist with the Beijing Tongren Hospital, Capital Medical University. Her research interests include glaucoma related surgeries, biomechanics, and tissue biology.

GUI-BIN BIAN (Member, IEEE) received the bachelor's degree in mechanical engineering from the North China University of Technology, Beijing, China, in 2004, and the master's and Ph.D. degrees in mechanical engineering from the Beijing Institute of Technology, Beijing, in 2007 and 2010, respectively. He is currently a Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interest includes design, sensing, and control for medical robotics.