# Deep Learning-based Intraoperative Video Analysis for Cataract Surgery Instrument Identification

Z. Guo*, Y.H. Chan and N.F. Law

EEE Dept, The Hong Kong Polytechnic University, Hong Kong SAR

E-mail: guozhe9922@gmail.com Tel: +86-15101112899

***Abstract* — Surgical instrument detection and classification is a critical task for enhancing surgical procedures monitoring, assisting surgical operations, supporting medical education, and enabling the development of intelligent surgical systems. However, there are a few challenges in this domain. The foremost concern is the impact of varying background conditions. Additionally, class imbalance presents another challenge, potentially leading to biased classification results. To solve these challenges, this study proposes a deep learning-based system consisting of two key components: an attention region detection module and a ResNet50 classification model. The attention region detection employs an optical flow-based method to incorporate both temporal and spatial information from the surgical video so that critical attention regions covering surgical instruments are identified. Our experimental results show that the classification accuracy can be improved from 58.7% to 81.9% by using the attention region detection component. To deal with the challenge of class imbalance, we use focal loss and interleaved sampling strategy as solutions. Interleaved sampling uses both the spatial and temporal information of surgical videos to balance the number of samples across different instrument classes, through which some scarce surgical instrument classes are expanded, thus preventing biased learning of the model. And the validation accuracy on the balanced dataset achieves 87.1%. This study demonstrates the effectiveness of deep learning techniques in addressing challenges in cataract surgery video analysis.**

## I. INTRODUCTION

Deep learning has made a huge contribution to today's medical field specifically in the fields of image recognition and image processing[1]. Cataract is a disease that is highly prevalent among the elderly. In recent years, deep learning technology has played an irreplaceable role in cataract surgery.

In the field of cataract surgery, the current research status of deep learning technology shows that its application in surgical procedures has potential and indeed has made certain progress[2][3]. These developments are of great significance. On one hand, by analyzing surgical videos and patient health data, deep learning models can not only provide real-time feedback to help doctors adjust treatment strategies and improve surgical success rates[3], but also predict potential complications after surgery[4][5][6] by analyzing clinical data and imaging, providing doctors with more comprehensive preoperative assessment and patient management strategies.

This study describes a system for identifying and classifying different surgical instruments used during cataract surgery. The core component of our system is a module that can determine the region of attention based on an optical flow method. With attention regions of surgical instruments as the input to the classification model, we can minimize the negative impact of the background portion. This study also proposes an interleaved sampling strategy to address the problem of class imbalance, through which the classification model can fairly learn the features of each class and effectively improve the accuracy.

## II. LITERATURE SURVEY

The identification and classification of cataract surgery stages and instruments have extraordinary significance in both medical surgery and artificial intelligence fields[1][7]. For example, [8] first segments parts of the pupil to extract the attention region and subsequently recognizes surgical instruments with that region using KNN algorithm. However, one of the challenges is that the attention region is not always accurately positioned at the center of the pupil and another is that when some interferences are present in the surgical video because of factors like pupil's outline distortion.

[9] addresses the challenge of domain shift between different dataset for surgical instrument recognition. Domain shift can be caused by variations in tools, video resolution, and other factors, leading to poor performance of trained models on

different dataset. To mitigate this issue, the authors proposed the unsupervised domain adaptation (UDA) method to improve the accuracy in cross-dataset scenarios. However, despite the improvement of the UDA method, the classification model generalizes poorly when faced with switching dataset. This point is further illustrated in [10]. The effect of the background portion of the surgical videos on classification accuracy was similarly illustrated in [8] [9] [10]. Therefore, it is important to extract the correct attention region to minimize the negative influences of the background in cataract surgery instrument classification.

[11] focuses on phase recognition in cataract surgery videos through deep learning techniques. The authors mention certain limitations, such as the difficulty in accurately predicting categories with infrequent occurrences and scarcity of data. These categories have very low recall, indicating that the model still needs more enhancements to handle less frequent occurrences. Therefore, enhancements are required to increase the size of the dataset, especially for the scarce categories. Zisimopoulos et al. raised the same problem of scarcity of samples of some categories in the dataset in [2].

## III. PORPOSED SURGICAL INSTRUMENT CLASSIFICATION SYSTEM

It should be noted that since our research in this study is a joint project with the Department of Ophthalmology of the United Hospital in Hong Kong, the 53 cataract surgery video dataset we used in the experiment were all provided by doctors from the hospital.

### *A. System Overview*

Fig.1 shows the block diagram of the system proposed in this study. The input to the system is intraoperative video clips. The video clip is sampled to obtain video frame pictures of size 1920×1080 each. These images are then used as input to the Attention Region Detection (ARD) Module. The ARD Module detects the area of interest in each frame through an optical flow method and then extracts the attention region in the frame. The size of the attention region is fixed to be 512x512 for all frames. Attention regions are then fed into the ResNet50 model for surgical instrument detection and classification. The ResNet50 model was trained with a dedicated training dataset for classifying one surgical instrument in an image of size 512×512.

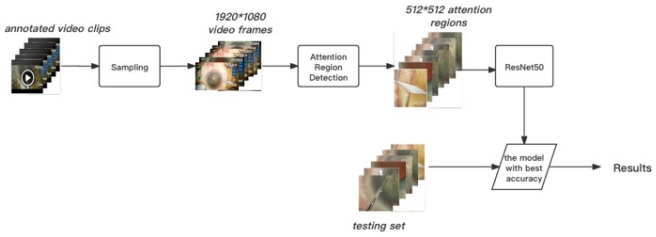


Fig.1 overview block diagram of proposed system

To eliminate the negative impact of these factors on the classification process, we introduced the Attention Region Detection Module to confine the condition of the input to the ResNet50 model. This Attention Region Detection Module uses the optical flow method to extract a 512×512 region that covers only one single surgical instrument within a raw 1920×1080 video frame image. Since the input is now well-conditioned and no downsampling was performed, the spatial features of the surgical instrument are well preserved during the preparation of the input. Consequently, the system is able to improve the classification accuracy remarkably.

### *B. Attention Region Detection*

In the proposed system, an attention region means the region that the exploited deep-learning model should put its focus on. The model should ignore other irrelevant regions when processing an image. The function of the ARD module is to intercept partial images that contain surgical instruments from the original video clips.

We assume that the surgical instruments appearing in a video frame are the foreground of the frame and all other parts are the background. Fig. 2 provides an example of separating the foreground and the background of a video frame. It is obvious that the background includes the patient's pupils and the parts around the eyes. Factors such as lighting and shooting angles bring a lot of uncertainty to the background. The foreground is actually our concern and it should be the key object in the attention region. Without loss of generality, we can assume that the background does not change too much within a very short time duration during the same surgical phase, while the foreground can have more variability in its position. By taking advantage of the characteristic that the background does not change in a short period of time, we use the optical flow method to determine the position of the surgical instrument, which is also the position of the foreground. More specifically, the

optical flow method compares the grayscale values of the current and the previous video frames to obtain the point of maximum change. This point is considered as the center of the 512×512 attention region of the current frame. The attention regions generated by the ARD module will be used as input to the classification and recognition model. If the target instrument is successfully captured in the attention region, then we can minimize the impact of the uncertain factors in the background on the classification results and make the classification more robust to the background in the video.

Fig. 2 an example of separating the background and the foreground of a video frame

To better separate the foreground and the background in a video frame, it should be noted that an appropriate sampling rate needs to be used when sampling video clips into video frame pictures. If the sampling interval is too short, the "foreground" may not move significantly, and the "foreground" cannot be successfully extracted. If the sampling interval is too long, the "background" may also undergo relatively drastic changes during this period.

*C. Methodologies*

i. Farneback Method

In this experiment, the signal received by the computer is often two-dimensional picture information. Optical flow field is a term used to describe the motion vector field of pixels reflected by moving objects in three-dimensional space in a two-dimensional image. In order to better locate the position of surgical instruments in video frames without being affected by the background, we used the Farneback[12] method. Naturally, for each pair of consecutive images in the dataset, we can compare the grayscale values between adjacent frames to find the point with the maximum difference in grayscale values. Subsequently, this point serves as the center for cropping the original image into a 512×512 pixel-sized picture. In this process, by choosing the appropriate sampling rate, adjusting the parameters of the optical flow method, etc., it is possible to make the cropped image to contain a localized image of the surgical instrument.

ii. Loss Function

Due to the significant differences in the duration of different surgical phases, the appearance times of different surgical instruments in the video also vary greatly. Considering the problem of class imbalance, we chose to use focal loss[13] as the loss function. Focal loss allows the network to focus more on difficult-to-classify samples, effectively improving the model's performance on minority samples.

*D. Results*

i. Optical Flow

Taking a set of adjacent pictures in Fig. 3 as an example, we calculate the optical flow of them. Then we use the Farneback method to determine the maximum change point between two adjacent frames. Fig. 4 shows the energy map. The scale bar on the right shows the energy value represented by each color in the figure. The higher the pixel color value, the greater the energy, indicating a greater change as determined by the optical flow method.

Similarly, according to the optical flow, the arrow map of the optical flow changes can be obtained as shown in Fig. 5. The direction pointed by the green arrow is the direction of motion of the pixel, and the length of the arrow represents the size of the motion vector of the pixel.

It can be seen from the energy map and arrow map that the Farneback method detects changes in the position of the surgical instruments in the adjacent legends mentioned above, and the changes in the background are much smaller than the changes in the surgical instruments.

ii. Attention Region Detection

After detecting the maximum change point in two adjacent pictures according to the Farneback method, a square of size 512×512 is cut out from the picture with this point as the center. Fig. 6 shows the square area intercepted with the maximum change point as the center.

iii. Model Performance

To mitigate the influence of uncertain factors in the “background”, it is necessary to extract the attention regions that contain surgical instruments separately from the video frames. Use pictures extracted by ARD module as input to ResNet50. In the ResNet50 model, in addition to using Focal Loss as the loss function to reduce the negative impact of uneven sample distribution, we also use dropout to prevent

over-fitting. After the experiment, we got the results in Fig. 7.

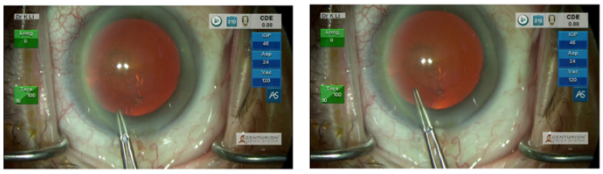

Fig. 3. a set of adjacent pictures in the dataset

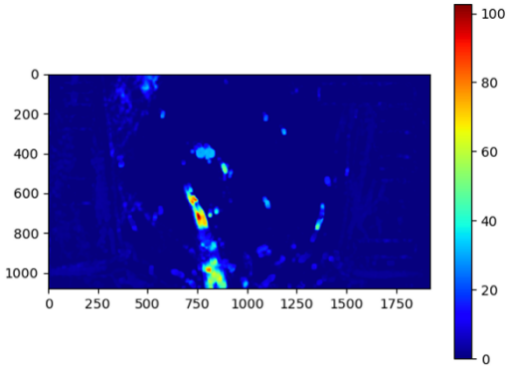


Fig. 4 energy map obtained based on optical flow

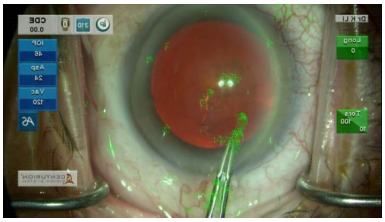

Fig. 5 arrow map obtained based on optical flow

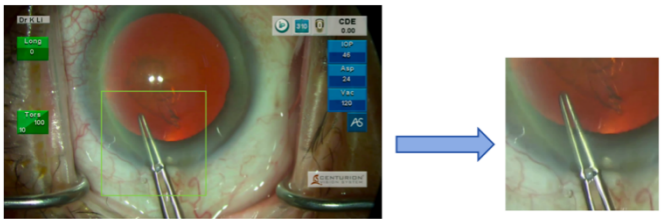

Fig. 6 attention region obtained based on maximum change point

According to Fig. 7, the accuracy of the best model reached 81.9%, but there is a difference between the accuracy of the validation set and the accuracy of the training set. In the line chart of various evaluation indicators, we can see that except for the early stage of training, the performance of the model in each epoch slowly improves without major fluctuations. This shows that the model maintains good stability while ensuring good training accuracy.

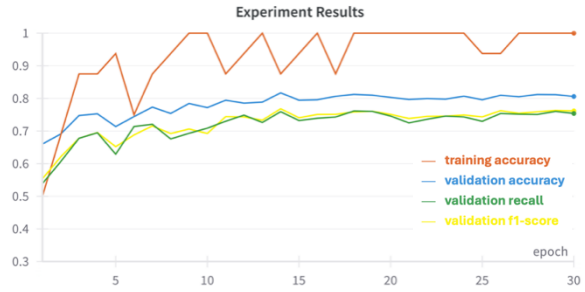


Fig. 7 training accuracy, validation accuracy, validation recall and f1-score

## *E. Ablation Study*

Before verifying the model performance, we performed ablation experiments. First, the performance of the system with only the ResNet50 model is tested, in which Cross Entropy is used as the loss function, and other parameters and settings remain the same.

After sampling the video frames at a sampling rate of 0.2 seconds per frame, they are directly put into the ResNet50. Based on the above steps, we obtained the following results after training:

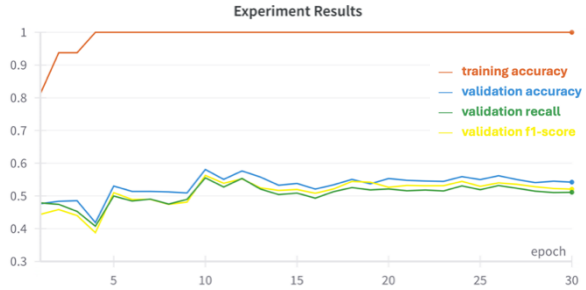


Fig. 8 training accuracy, validation accuracy, validation recall and f1-score

The best model we obtained in this ResNet50 model experiment achieved an accuracy of 58.1% on the validation set. It can be found that the accuracy on the training set reaches 100%, while the accuracy on the validation set is only 50%-60% on average. We believe there may be three reasons. First, the over-fitting phenomenon is serious during the training process. Second, because the data of the training set and validation set are sampled from different surgical videos, differences between different videos, such as shooting angles, lighting factors, patient's eye status, etc., lead to differences in the performance of the model on the training set and validation set. The third is that the imbalance of sample distribution causes the accuracy of the validation set in this experiment to be much lower than

the accuracy of the training set. Therefore, we replaced the originally used Cross Entropy loss function with Focal Loss to reduce the impact of uneven sample distribution on classification accuracy. Fig. 9 shows the experiment result.

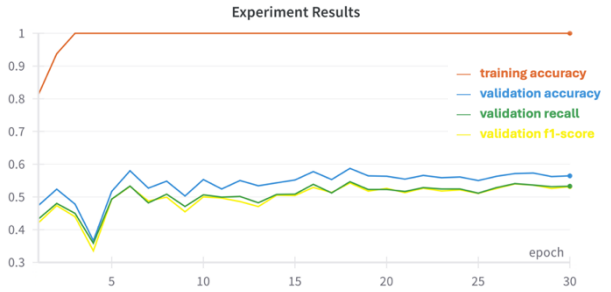


Fig. 9 training accuracy, validation accuracy, validation recall and f1-score

Compared with Fig. 8, after using Focal Loss, the best accuracy of the validation set increased from 58.1% to 58.7%. Overall, there is no significant improvement in model performance. Therefore, after the ablation experiments, we focused on solving three issues that would arise from speculation. The last chapter has proven that the "background" difference of different videos is too large and the over-fitting phenomenon does have a negative impact on the recognition of the model.

## IV. PRODUCTION OF BALANCED TRAINING DATASET

### *A. Dataset*

In order to avoid the problem of model "cheating", which occurs when frames from the same video are present in both the training and testing dataset, leading to the model performing exceptionally well, we believe that the images in the training set, validation set, and testing set should be generated from different videos. Therefore, in order to have a fairly unbiased dataset to train the model and test the performance of the model, a preliminary division was made: among these 53 videos, 30 were used for the training set, 15 were used for the validation set, and 8 were used for the testing set, and the videos used between them did not overlap with each other.

### *B. Challenge*

After detailed annotation and sampling of the dataset, Fig. 10 shows the sample number distribution of all classes:


Fig 10. Quantity distributions of different classes

It can be seen that the problem of unbalanced sample distribution is very serious. This may cause the model to over-learn features for classes with a large number of samples and not be able to fully and thoroughly learn features for scarce samples.

### *C. Interleaved Sampling*

In order to verify whether the balance of sample data of each class in the dataset will affect the final classification accuracy, we use the interleaved sampling operation to expand the scarce categories of surgical instruments, so that the number of samples of each category is balanced. Figure 4.2 shows the principle of interleaved sampling. Suppose there are N consecutive video frames, and we take one frame every 6 frames. Then we can take a series of frame pictures every 6 frames such as Frame 1, Frame 7, Frame 13, Frame 19, Frame 25, etc. as one set. Similarly, we can also use any one of the second to the sixth frames as the starting frames to extract different video frame groups. Each group can be used as input to the Attention Region Detection Module to find the maximum change point between two adjacent frames to provide input data for the classification model. And the video frames in frame groups extracted using the picture frame before the seventh frame as the starting frame do not overlap each other.

The significance of interleaved sampling is obvious. By utilizing both temporal and spatial information, the amount of data can be increased almost exponentially, while the extracted data is not repeated. The temporal information of video frames provides a basis for the arrangement and extraction of video frames, while the spatial information not only allows all video frames to be fully utilized but also allows video frames to be divided into groups and extracted without being reused.

### *D. Experiments*

In order to explore whether the problem of unbalanced sample distribution has a negative impact on the model, we performed an interleaved sampling operation on the dataset

used in the previous two parts, which can expand the dataset and make the distribution of samples of each class balanced. Fig. 12 is the sample distribution of each class after interleaved sampling operation.

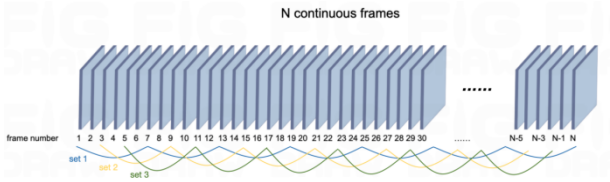


Fig. 11 N continuous frames using interleaved sampling


Fig. 12 quantity distribution of different classes after interleaved sampling

Taking this balanced dataset as the input of the system, and keeping other conditions all the same, we obtained the following result in Fig. 13, in which we can see the accuracy of the best model on the verification set reached 87.1%, and all indicators tended to be stable throughout the training process. This shows that a balanced distribution of samples in each class contributes positively to the accuracy of the classification task. After ablation study and experiments conducted with ARD module and the balanced dataset, we got the results in Table. 1, in which we can see that the effectiveness of ARD module in determining attention region. And it also proves that a balanced dataset has a benign effect on model learning.

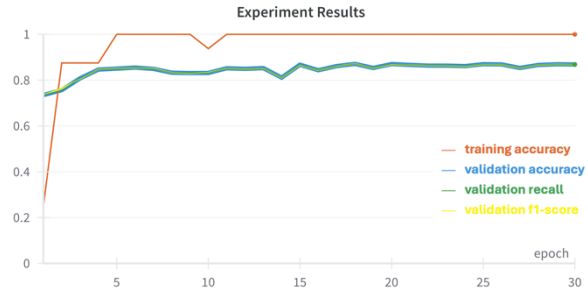


Fig. 13 training accuracy, validation accuracy, validation recall and f1-score

Table. 2 shows the performance comparison between our proposed model and the ResNet50 model. The compared indicators include accuracy, precision, recall, and f1-score. Their inputs are identical video clips. It should be noted that they perform different preprocessing on video clips. Our model uses the Farneback method to find the maximum change point and then crops the image to a size of 512×512. ResNet50 model samples video clips at a sampling rate of 5 frames per

Table. 1 The validation accuracy of different methods used.

| | Version1 | Version2 | Version3 | Version4 |
|---|---|---|---|---|
| Focal Loss | × | √ | √ | √ |
| Dropout | × | × | √ | √ |
| Attention Region Detection | × | × | √ | √ |
| Interleaved Sampling | × | × | × | √ |
| Accuracy | 58.1% | 58.7% | 81.9% | 87.1% |

Table. 2 Comparison withResNet50 model. The values used for precision, recall and F1-score are all macro averages.

| Method | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| ResNet50 | 0.594 | 0.642 | 0.594 | 0.588 |
| **Ours** | **0.871** | **0.872** | **0.868** | **0.868** |

second, and then perform resize operations to improve the efficiency of training and inference. It resizes the video frame to 225×225 and then center-crop it to 224×224.

## V. CONCLUSIONS

This study addresses the key challenges associated with surgical instrument detection and classification in cataract surgery videos. The proposed deep learning-based system contains two key components, namely an Attention Region Detection Module and a ResNet50 classification model. The former incorporates both temporal and spatial information to extract attention regions while the latter focus on these attention regions and performs surgery instrument classification. The Attention Region Detection Module based on the Farneback method plays an indispensable and significant role in promoting this research. It can well extract critical regions of surgical instruments from video clips, which minimizes the negative impact of uncertainties in background portion on the model's ability to accurately identify surgical instruments. As for class imbalance issue, the interleaved sampling strategy effectively utilizes temporal information and spatial information to better solve the problem of uneven distribution of sample numbers. Class balancing can significantly improve the generalization ability and fairness of the model, ensuring that the model is not biased toward any specific class.

## REFERENCES


[1] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1), 53. DOI: 10.1186/s40537-021000444-8.

[2] Zisimopoulos, O. et al. "DeepPhase: Surgical Phase Recognition in CATARACTS Videos." Medical Image Computing and Computer Assisted Intervention – MICCAI 2018.

[3] Lee, M., Kang, J., Lee, S., Kim, H., & Lee, K. (2023). Real-time estimation of the remaining surgery duration for cataract surgery using deep convolutional neural networks and long short-term memory. Translational Vision Science & Technology, 12(2), 24.

[4] Quellec, G., Lamard, M., Cochener, B., & Cazuguel, G. (2014). Real-Time Segmentation and Recognition of Surgical Tasks in Cataract Surgery Videos. IEEE Transactions on Medical Imaging, 33(12), 2352-2360.

[5] Cheng, B., Li, H., & Qu, X. (2020). Deep Learning-Based Keratoconus Screening Using Corneal Topographic Maps. Investigative Ophthalmology & Visual Science, 61(10), 45.

[6] Abbas, A., Waqar, S., Makary, M., et al. (2019). Machine Learning for Prediction of Postoperative Complications after Hepato-Biliary and Pancreatic Surgery. Annals of Surgery, 270(1), 117-125.

[7] N. Ghamsarian, M. Taschwer, D. Putzgruber-Adamitsch, S. Sarny and K. Schoeffmann, "Relevance Detection in Cataract Surgery Videos by Spatio- Temporal Action Localization," *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 10720-10727, doi: 10.1109/ICPR48806.2021.9412525.

[8] Palepu, V., Sundar, R., Reddy, A. R., Manjunath, M. K., & Yegnanarayana, B. (2019). Surgical tools recognition and pupil segmentation for cataract surgical process modeling. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (pp. 1730-1734). IEEE. doi:10.1109/ISBI.2019.8759429

[9] Pasa, F., Deepak, R., Bria, A., Nappi, M., & Ricci, E. (2020). Cross-dataset adaptation for instrument classification in cataract surgery videos. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020 (pp. 620-628). Springer, Cham. doi:10.1007/978-3-030-59716-0_60

[10] Sokolova, N., Schoeffmann, K., Taschwer, M., Putzgruber-Adamitsch, D., & El-Shabrawi, Y. (2020). Evaluating the Generalization Performance of Instrument Classification in Cataract Surgery Videos. In Proceedings of the 26th International Conference on Multimedia Modeling (MMM2020) (pp. 626-636). Springer, Cham. DOI: 10.1007/978-3-030-37734-2_52.

[11] Hsu-Hang Yeh, Anjal M Jain, Olivia Fox, Sophia Y Wang. "PhacoTrainer: Deep Learning for Activity Recognition in Cataract Surgical Videos." Investigative Ophthalmology & Visual Science, vol. 62, no. 8, 2021, p. 583. DOI: 10.1167/iovs.62.8.583.

[12] Gunnar Farneback "Two-Frame Motion Estimation Based on Polynomial Expansion," Scandinavian Conference on Image Analysis (SCIA 2003), pp. 363-370, (2003).

[13] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2980-2988).