Human eye iris and pupil segmentation using infrared camera snapshots

A. Samarin[1], A. Toropov[1], A. Savelev[1], A. Dzestelova[1], A. Nazarenko[1], A. Motyko[2], A. Golovatiuk[1], P. Dmitriev[1], E. Mikhailova[1], and V. Malykh[1]

[1] ITMO University, St. Petersburg, 197101 Russia
[2] St. Petersburg Electrotechnical University "LETI", St. Petersburg, 197022 Russia

Abstract. At the present time, problems related to the automation of medical data processing are urgent. Particularly systems for human physiological parameters monitoring and analysis are under great attention. Such systems often use special types of sensors such as infrared cameras for biomedical images capturing. The current article describes our research on human pupil and iris segmentation problem over snapshots acquired by infrared camera. In the article we propose a custom DNN architecture with novel loss function for training human segmentation eye. We also present the segmentation dataset and comparative analysis of various segmentation technique applied to human eye segmentation problem. The proposed model outperforms other considered approaches and demonstrate the state-of-the-art results.

Keywords: Image segmentation · Image processing · Biomedical image analysis.

1 Introduction

At the present time, the automation of medical processes and its applications related to statistical data analysis and machine learning are under active development [27, 26].

Among the wide variety of issues dealing with processing of raw images, problems related to the analysis of human eye can be highlighted. The analysis of infrared camera snapshots is actively used to evaluate the physiological parameters of the human eye. Thus, the problem of infrared human eye segmentation has great industrial and scientific-medical applications. For example, a system of psychophysiological assessment (under development by NPP VIDEOMIX [3]) uses the results of segmentation of infrared snapshots of the human eye.

Nowadays, there are many approaches to the segmentation of objects in the image [2, 13, 24]. These approaches can be conditionally divided into several groups. First group is a classical methods represented by an adaptation of cluster analysis approaches to highlight segments in the images [2, 13]. The second group is represented by segmentation methods based on deep neural network

[3] NPP VIDEOMIX - https://v-mix.ru

architectures [12, 1, 18]. And the third group includes approach that we have developed are presented with special biomedical image segmentation engines [24, 20, 27, 33, 4, 5].

However, the specific context of the human eye image segmentation task based on infrared camera images shows insufficient efficiency of the aforementioned approaches in the context of their use in such complexes as MIX GT-19 [4]. This circumstance led us to develop our own methods for the task of human eye segmentation in such conditions.

Thus our contribution is presented by the following aspects:

- we publish an annotated infrared snapshots dataset for segmentation of human iris and pupil (Iris and Pupil dataset - InP [5]).
- we present a comparative analysis of classical methods and neural network methods for segmentation of human iris and pupil images.
- we present our method that demonstrates the state-of-the-art on the proposed dataset.
- we present a comparative analysis of different image preprocessing methods and design approaches for building deep neural network architectures to solve our problem.
- we also propose an additional segmentation quality metric to evaluate the quality of the segmentation of the eye elements (smoothness estimation).

2 Related work

'Classical' methods for human eye segmentation involve adapting cluster analysis approaches to highlight segments in human eye images. These methods typically involve grouping pixels in the image into clusters based on their similarities and differences in intensity or color. By analyzing the resulting clusters, these methods can identify regions of the image that are likely to correspond to the iris, pupil, and other parts of the eye.

One popular approach in this category is the k-means clustering algorithm, which partitions the image pixels into a predetermined number of clusters based on their euclidean distance from a set of randomly initialized cluster centroids [15, 23]. The k-means algorithm offers various benefits such as simplicity and efficiency in processing large amounts of data. To apply k-means to segmentation specifically, researchers have used features like intensity, texture, and shape for clustering. For example, we can encode an image using Gabor wavelets and then apply k-means to group similar pixels together [22, 28]. Alternatively, we can use a hybrid approach that combines k-means with other techniques like morphological operations [14] or graph-based methods [6]. However, like any other clustering algorithm, k-means does not always produce optimal results, especially when dealing with images that have complex structures or low quality.

[4] MIX GT-19 - https://v-mix.ru/technology/eye-tracking/
[5] Iris and Pupil dataset (InP) - https://github.com/itmo-cv-lab/eye-tracking-dataset

Another common clustering method, mean-shift, can be utilized for grouping pixels with similar characteristics, such as color or texture, into regions of interest. One advantage of mean-shift over other clustering algorithms like k-means is its ability to handle non-parametric distributions and to automatically determine the number of clusters.

Fukunaga et all [9] introduced the mean shift procedure based on asymptotic unbiasedness, consistency and uniform consistency of a nonparametric density function gradient estimate using a generalized kernel approach. The mean shift procedure has found its applications in image analysis [17] and texture segmentation, among other fields.

Despite its advantages, mean-shift can be computationally expensive and may take longer to run than other methods.

While mentioned above 'classical' methods have been widely used for eye detection in the past, they often struggle with variations in scale, orientation, and lighting conditions. As a result, more modern deep learning-based approaches have emerged that can handle these challenges more effectively.

The popularity of *Convolutional Neural Networks (CNN)* has increased due to their ability to generalize well and the availability of powerful graphics processing units that can efficiently solve optimization problems with a large number of parameters. The first work on iris and eye segmentation using CNN appeared in the early 2016s [7, 8]. One of the first successful approaches was the DeepIris [10] network architecture, which used convolutional and pooling layers to extract features from iris images. The extracted features were then transferred to a fully connected block that produces a segmentation result.

In the following years, many modifications of CNN architectures for iris and eye segmentation were proposed. For example, various adaptations of UNet [25], a convolutional neural network designed specifically for segmenting biomedical images, have been utilized for the segmentation of eye structures. The UNet architecture utilizes convolutional network concept along with downsampling, upsampling, and bottlenecks. Another architecture, ENet [19], has also implemented a encoder-decoder structure with similar bottlenecks to diminish model complexity for real-time segmentation in mobile applications. In particular, recent studies investigates MinENet [21], EyeNet [16] and EyeMMS [3] architectures, which are based on ENet and UNet. These models were tested using OpenEDS dataset [11] consisting of 12,759 pixel-level annotated images for key eye regions, including iris, pupil, and sclera, as well as 252,690 unlabeled eye images. The results showed values over 0.92 on the mIOU metric, indicating high performance in iris and pupil segmentation. However specific context proposed by our infrared snapshots dataset (InP) that was prepared in cooperation with NPP VIDEOMIX, does not allow to obtain sufficient performance in segmentation for MIX GT-19 eye-tracking system integration.

There is an infrared human eye segmentation approach described in [31]. Unfortunately special shooting conditions do not allow to apply the proposed solution in the context of a general perspective problem, which we consider in our work.

3 Problem statement

Thus the goal of the article is to describe our research results on human eye segmentation over infrared camera full face portrait snapshots. Thus, our focus is on the problem of segmenting a person's eye in a portrait photograph, which can be summarized as follows.

We use a monochrome image I captured using an infrared camera as an input. As a segmentation result we have to obtain binary masks of human eyes pupil (M_{pupil}) and iris (M_{iris}). Output binary masks that present human eyes segmentation result can be described as matrices which elements can be assigned with only two values according to the following. $M_{pupil}[x, y] = 1$, if pixel with $[x, y]$ coordinates refers to the area occupied by the pupil, else 0, and $M_{iris}[x, y] = 1$, if pixel with $[x, y]$ coordinates refers to the area occupied by the iris, else 0.

4 Proposed solution

To solve the problem we used a custom neural network architecture, which is based on the Unet-like convolutional neural network model with an attention mechanism. Our approach is similar to [27, 26] attention integration but uses multi-scale self-attention that description we provide below. We based our approach on Unet-like architecture because Unet-like architectures provide good results in biomedical images segmentation. We also used multi-scale self-attention as a special non-local block as it allows to take into consideration non-local image patterns.

So the solution of the stated problem is presented with a combined neural network that pipeline contains the following stages: preprocessing and segmentation. When training the neural network architecture, we took into account the characteristics of the object, namely the smoothness and convexity of the contours. Let us have a closer look at the model components.

4.1 Data preprocessing

For better quality of image segmentation, we used the preprocessing stage. There are a lot of solutions for an automatic general images enhancement like [30, 29]. However our domain is very specific so we have to combine our custom preprocessing pipeline that is based on the following filters usage. In order to obtain the most optimal combination of transformations of the analysed image, we used various combinations and parameterisations of different filters and selected successive application of described below transformations.

Median filtering A median filter is an image processing filter used to remove noise from an image. It works by replacing each pixel with the median kernel value (a square area defined by the kernel size) surrounding that pixel. This reduces the effect of outliers (some pixels that have values that are significantly different from neighbouring pixels).

Gaussian blur filtering The Gaussian blur filter represents a local kernel convolution, defined by samples from a two-dimensional Gaussian function. This function is the product of two one-dimensional Gaussian functions:

$$G_{\sigma,\mu_x,\mu_y}(x,y) = \frac{1}{(2\pi\sigma^2)} exp(\frac{-((x-\mu_x)^2 + (y-\mu_y)^2)}{2\sigma^2}),$$

where μ_x, μ_y - the mathematical expectations of the x- and y-axes, σ - standard deviation (σ^2 - variance), which is also called the radius of this function, e - Euler's number. This function is the product of two one-dimensional Gaussian functions.

Centred Gaussian filtering If we assume that the Gaussian function is centred (i.e. mean $\mu_x = \mu_y = 0$), the formula is simplified:

$$G_\sigma(x,y) = \frac{1}{(2\pi\sigma^2)} exp(\frac{-(x^2+y^2)}{2\sigma^2}).$$

Such a centred function is computed at (2k + 1)×(2k + 1) points, with the initial window pixel at (0, 0). This results in an important filter kernel for local operator with parameters $\sigma > 0$ and $k \geq 1$.

Gaussian pyramid filtering The Gaussian pyramid is a multiple anti-aliasing and shrinking of the image to a certain level to give smaller versions of the image at different sizes, but with the key details.

Sigma filtering This local operator is also defined for windows $W_p(I)$ of size (2k + 1)×(2k + 1) at $k \geq 1$. The parameter $\sigma > 0$ is interpreted as an approximation of the noise accompanying the image I (e.g., σ is approximately 50 if $G_{max} = 255$). In the proposition that the local operator is computed in parallel, the new image J is generated as follows:

1. Compute the window histogram $W_p(I)$;
2. Calculate the average μ of all values in the interval [I(p) - σ, I(p) + σ];
3. Put J(p) = μ.

Sharpness filtering The purpose of sharpening is to produce an improved image J by increasing the contrast of the original image I along the edges without adding noise to the homogeneous areas. This local operator calculates the difference $R(p) = I(p) - S(p)$ between the original and smoothed images. The difference is then added to the original image I:

$$J(p) = I(p) + \lambda[I(p) - S(p)] = [1 + \lambda]I(p) - \lambda S(p),$$

where, $\lambda > 0$ is the scaling factor. In principle, to obtain a smoothed image $S(p)$.

The dimensional parameter k ("radius") of these operators controls the spatial distribution of the smoothing effect, and the parameter λ controls the effect

of the correction signal $I(p) - S(p)$ on the final result. Thus, k and λ are the usual interactive control parameters for unsharp masking.

The best configuration is to use the Gaussian pyramid blur to remove noise from the image, followed by the addition of a sharpening filter to emphasise the sharpness of the edges.

4.2 Deep neural network architectures

Our architecture is based on the UNet deep neural network model and the method proposed in this article [27]. But unlike approach proposed in the paper [27], we used attention at all scales and one output presented with a binary mask (Fig. 1).

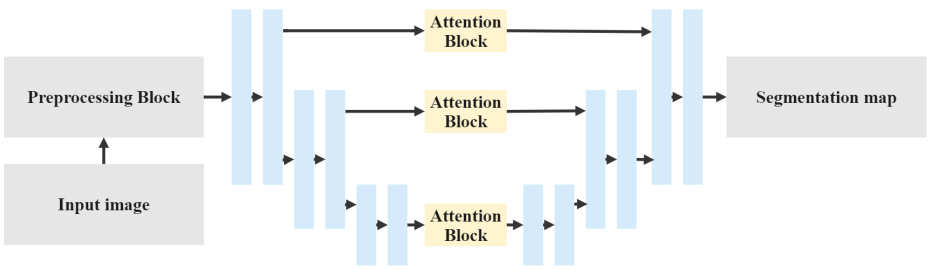

Fig. 1. The proposed DNN architecture.

As blocks of attention, we used the self-attention mechanism that was used in [27, 26]. Our self-attention block (Fig. 1) is combined with convolutional UNet features extractor at different scales.

4.3 Special loss function

It should be noted that objects under consideration have very specific properties. In particular, appropriate segmentation (Fig. 3) results have the rounded shape and also locally certain properties of curvature or concavity. That is conditioned by anatomical features of a structure of a human eye. In order to take this feature into account as much as possible during our segmentation model training, we have developed a special additional loss function component.

The main idea is to divide the segmentation result into sections and calculate the statistics of the local properties of convexity and concavity. After that, statistics are collected on local properties for each patch and a penalty is added

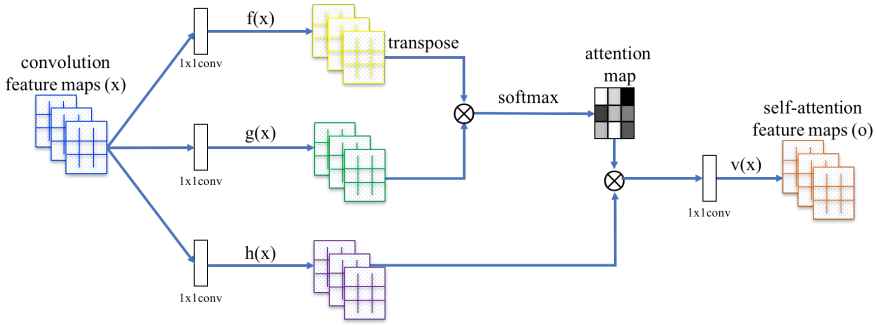

Fig. 2. Self-Attention block scheme. Image adopted from [32].

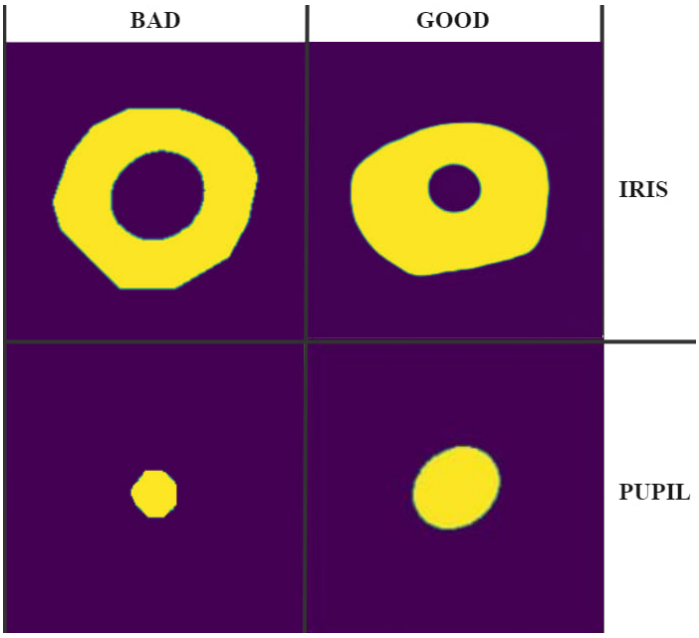

Fig. 3. Examples of different quality segmentation results.

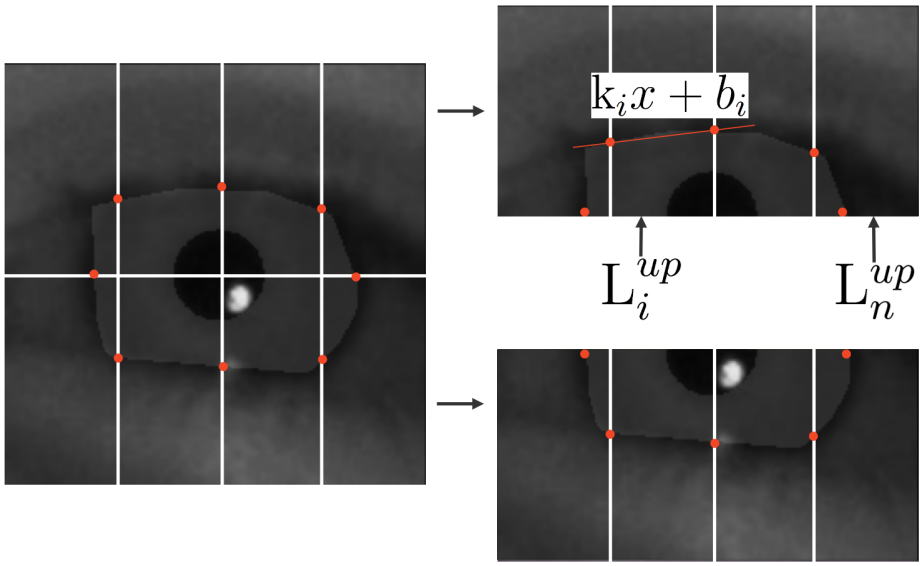

Fig. 4. Special loss function evaluation illustration.

for the deviation from the expected properties of an object having a spherical shape (see Fig. 4).

Let us describe in more detail the algorithm for calculating the special component of the loss function L_{sp}

Algorithm 1 An algorithm of L_{sp} evaluation

1: calculate coordinates of centroid of segmentation result M
2: split M into two parts L_{up} and L_{down}
3: split M_{up} and M_{down} into n regions of the same area
4: evaluate penalty

$$L_{sp} = L_{up} + L_{down} = \frac{1}{n} * \sum_{i=1}^{n} (l_i^{up} + l_i^{down}),$$

where $l_{iup} = 1$ if $\sum L_i^{up}[x_i^{above}, y_i^{above}] < threshold^{up} * \sum L_i^{up}[x_i^{below}, y_i^{below}]$, $y_i^{above} < k_i^{up} * x_i^{above} + b_i^{up}$, $y_i^{below} > k_i^{up} * x_i^{below} + b_i^{up}$, else 0 . l_{idown} computed similarly but using $threshold_{down}$.

5 Evaluation

We investigated performance of the proposed model and its variations using our dataset (InP).

5.1 Dataset

In our research, we have used a dataset (InP), which we have prepared together with NPP VIDEOMIX and made publicly available[6]. That dataset is represented with annotated pairs of the following form: an image of a human eye captured by an infrared camera, and a binary segmentation mask corresponding to a particular object (pupil and iris) (Fig. 5). The first patch of dataset contains 1758 images and related to iris segmentation. Infrared snapshots are paired with masks. Snapshots captured for 8 subjects (439 images per subject on average). The second part of the dataset related to pupil segmentation contains 7343 infrared snapshots of pupil paired with corresponding segmentation mask acquires from 21 subjects (699 images per subject on average). We used 5:1 train/test split ratio.

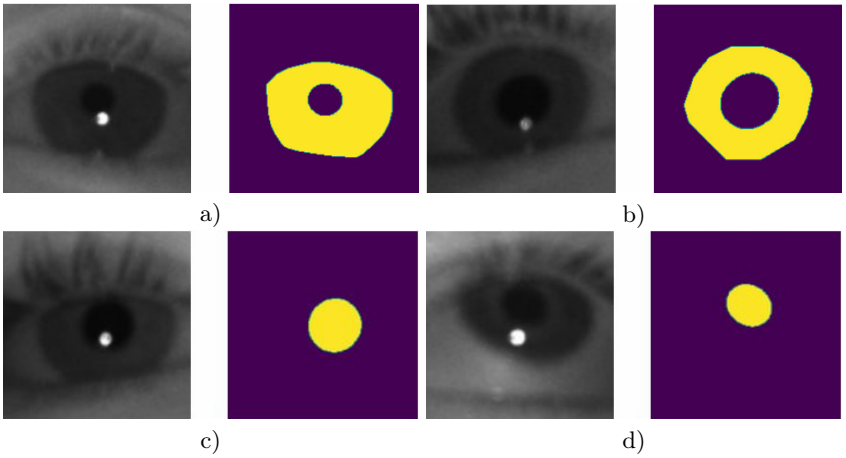

Fig. 5. *InP dataset illustration: a) iris and mask; b) iris and mask; c) pupil and mask; d) pupil and mask.*

5.2 Training details

We trained each model for approximately 30 epochs using early stopping. We determined the optimal batch size experimentally and found that a batch size of

[6] https://github.com/itmo-cv-lab/eye-tracking-dataset

32 provided the best performance. We trained our models using a cluster with multiple NVIDIA T4 16GB GPU workers. The performance results are presented in Section 5.3.

5.3 Experimental results

During our research, we employed a variety of practices to gain a better understanding of the problem at hand. Our findings were then organized and reported in Table 1. We used IoU as a segmentation quality measure.

Table 1. Comparative analysis of segmentation methods using the InP dataset

Model	**mIoU iris**	**mIoU pupil**
DeepIris	0.864	0.868
ENet	0.866	0.873
MinENet	0.872	0.881
EyeNet	0.879	0.884
EyeMMS	0.884	0.886
MinENet	0.888	0.891
Unet ResNet-18	0.872	0.882
Unet ResNet-50	0.897	0.901
Unet EfficientNet-b4	0.932	0.945
Unet++ EfficientNet-b4	0.936	0.951
Ours	**0.969**	**0.976**

As can be seen from the table 1, our approach outperforms counterparts and demonstrates state-of-the-art results on InP dataset.

6 Conclusion

In conclusion, our article described our research on human pupil and iris segmentation problem using snapshots captured by infrared camera. We proposed a custom architecture and novel special loss for training human eye segmentation neural network models. We also provided a human annotated segmentation dataset and compare analysis of segmentation technique approaches applied to the dataset. The experiment results demonstrate that our model outperforms other consideration approaches on the proposed dataset and achieved the state-of-the-art result.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. CoRR **abs/1511.00561** (2015), http://arxiv.org/abs/1511.00561

2. Bora, D.J., Gupta, A.K.: Clustering approach towards image segmentation: An analytical study. CoRR **abs/1407.8121** (2014), http://arxiv.org/abs/1407.8121
3. Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A.: Eye-mms: Miniature multi-scale segmentation network of key eye-regions in embedded applications. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
4. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. CoRR **abs/1606.00915** (2016), http://arxiv.org/abs/1606.00915
5. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017), http://arxiv.org/abs/1706.05587
6. Choong, M.Y., Khong, W.L., Kow, W.Y., Angeline, L., Teo, K.T.K.: Graph-based image segmentation using k-means clustering and normalised cuts. In: 2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks. pp. 307–312. IEEE (2012)
7. Fuhl, W., Santini, T., Kasneci, G., Kasneci, E.: Pupilnet: Convolutional neural networks for robust pupil detection. arXiv preprint arXiv:1601.04902 (2016)
8. Fuhl, W., Santini, T., Kasneci, G., Rosenstiel, W., Kasneci, E.: Pupilnet v2. 0: Convolutional neural networks for cpu based real time robust pupil detection. arXiv preprint arXiv:1711.00112 (2017)
9. Fukunaga, K., Hostetler, L.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on information theory **21**(1), 32–40 (1975)
10. Gangwar, A., Joshi, A.: Deepirisnet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition. In: 2016 IEEE international conference on image processing (ICIP). pp. 2301–2305. IEEE (2016)
11. Garbin, S.J., Shen, Y., Schuetz, I., Cavin, R., Hughes, G., Talathi, S.S.: Openeds: Open eye dataset. arXiv preprint arXiv:1905.03702 (2019)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. CoRR **abs/1703.06870** (2017), http://arxiv.org/abs/1703.06870
13. Isensee, F., Petersen, J., Kohl, S.A.A., Jäger, P.F., Maier-Hein, K.H.: nnu-net: Breaking the spell on successful medical image segmentation. CoRR **abs/1904.08128** (2019), http://arxiv.org/abs/1904.08128
14. Jardim, S., António, J., Mora, C.: Graphical image region extraction with k-means clustering and watershed. Journal of Imaging **8**(6), 163 (2022)
15. Jin, L., Xiao, F., Haopeng, W.: Iris image segmentation based on k-means cluster. In: 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems. vol. 3, pp. 194–198 (2010). https://doi.org/10.1109/ICICISYS.2010.5658566
16. Kansal, P., Devanathan, S.: Eyenet: Attention based convolutional encoder-decoder network for eye region segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3688–3693. IEEE (2019)
17. Liu, L.x., Tan, G.z., Sami Soliman, M.: Color image segmentation using mean shift and improved ant clustering. Journal of Central South University **19**(4), 1040–1048 (2012)
18. Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. CoRR **abs/1606.04797** (2016), http://arxiv.org/abs/1606.04797

19. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147 (2016)
20. Pathan, S., Tripathi, A.: Y-net: Biomedical image segmentation and clustering (2020)
21. Perry, J., Fernandez, A.: Minenet: A dilated cnn for semantic segmentation of eye features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
22. Premana, A., Wijaya, A., Soeleman, M.: Image segmentation using gabor filter and k-means clustering method. pp. 95–99 (10 2017). https://doi.org/10.1109/ISEMANTIC.2017.8251850
23. Qian, Z., Xu, D.: Automatic eye detection using intensity filtering and k-means clustering. Pattern Recognition Letters **31**(12), 1633–1640 (2010)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR **abs/1505.04597** (2015), http://arxiv.org/abs/1505.04597
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
26. Samarin, A., Savelev, A., Toropov, A., Dzestelova, A., Malykh, V., Mikhailova, E., Motyko, A.: One-staged attention-based neoplasms recognition method for single-channel monochrome computer tomography snapshots. Pattern Recognition and Image Analysis **32**(3), 645–650 (2022)
27. Samarin, A., Savelev, A., Malykh, V.: Two-staged self-attention based neural model for lung cancer recognition. In: 2020 Science and Artificial Intelligence conference (SAI ence). pp. 50–53. IEEE (2020)
28. Shi, H., Lee, W.L.: Image segmentation using k-means clustering, gabor filter and moving mesh method. The Imaging Science Journal **69**(5-8), 407–416 (2021)
29. Tatanov, O., Samarin, A.: Lfiem: Lightweight filter-based image enhancement model. 2020 25th International Conference on Pattern Recognition (ICPR) pp. 873–878 (2021)
30. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxim: Multi-axis mlp for image processing. CVPR (2022)
31. Valenzuela, A., Arellano, C., Tapia Farias, J.: Towards an efficient segmentation algorithm for near-infrared eyes images. IEEE Access **8**, 171598–171607 (01 2020). https://doi.org/10.1109/ACCESS.2020.3025195
32. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. Proceedings of Machine Learning Research, vol. 97, pp. 7354–7363. PMLR, Long Beach, California, USA (09–15 Jun 2019), http://proceedings.mlr.press/v97/zhang19d.html
33. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. CoRR **abs/1807.10165** (2018), http://arxiv.org/abs/1807.10165

Response to review

- "The authors use passive voice extensively." — **Fixed**.

- "Also, the Goal, Problem, and Solution of the paper are not explicitly mentioned." — **Fixed**.

- "Problem statement is not clear. The problem statement is one-sentence-paragraph that is hard-to-follow." — **Fixed**.

- "The authors provide a solution for the given challenge. However, they do not justify their solution (why exactly this solution is preferred)." — **Fixed**.

- "Unfortunately, some analogues are probably missed from consideration, e.g. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9200989" — **Fixed**.

- "The problem statement is clearly insufficient. It needs to be developed further." — **Fixed**.

- "The dataset used for testing needs to be explained with more detail. There is no indication of the number of instances used etc." — **Fixed**.

- "Table 1 caption needs to contain useful information." — **Fixed**.

- "English needs to be checked. As an example: "certain degree of convention" (??)" — **Fixed**.

- "rewrite the abstract." — **Fixed**.

- "change the keywords into something related to the conference." — **Fixed**.

- "Add caption to tables and figures." — **Fixed**.

- "Explain with detail the dataset used for the experiments." — **Fixed**.