Hindawi
Security and Communication Networks
Volume 2022, Article ID 1343766, 11 pages
https://doi.org/10.1155/2022/1343766

Research Article

Visual Attention and Motion Estimation-Based Video Retargeting for Medical Data Security

Qingfang Liu,[1,2] **Baosheng Kang**,[1] **Qiaozhi Hua**,[3] **Zheng Wen**,[4] **and Haipeng Li**[5]

[1]*School of Information Science and Technology, Northwest University, Xi'an 710127, China*
[2]*Network Center, Shijiazhuang Posts and Telecommunications Technical College, Shijiazhuang 050021, China*
[3]*Computer School, Hubei University of Arts and Science, Xiangyang 441000, China*
[4]*School of Fundamental Science and Engineering, Waseda University, Tokyo 1698050, Japan*
[5]*Capinfo Company Ltd, Beijing 100010, China*

Correspondence should be addressed to Qiaozhi Hua; 11722@hbuas.edu.cn

Received 15 January 2022; Accepted 12 July 2022; Published 9 August 2022

Academic Editor: G. Thippa Reddy

Medical data security is an important guarantee for intelligent medical system. Medical video data can help doctors understand the patients' condition. Medical video retargeting can greatly reduce the storage capacity of data on the premise of preserving the original content information as much as possible. The smaller volume of medical data can reduce the execution time of data encryption and threat detection algorithm and improve the performance of medical data security methods. The existing methods mainly focus on the temporal pixel relationship and foreground motion between adjacent frames, but these methods ignore the user's attention to the video content and the impact of background movement on retargeting, resulting in serious deformation of important content and area. To solve the above problems, this paper proposes an innovative video retargeting method, which is based on visual attention and motion estimation. Firstly, the visual attention map is obtained from eye tracking data, by K-means clustering method and Euclidean distance factor equation. Secondly, the motion estimation map is generated from both the foreground and background displacements, which are calculated based on the feature points and salient object positions between adjacent frames. Then, the visual attention map, the motion estimation map, and gradient map are fused to the importance map. Finally, video retargeting is performed by mesh deformation based on the importance map. Experiment on open datasets shows that the proposed method can protect important area and has a better effect on salient object flutter suppression.

1. Introduction

With the rapid development of high-tech medical imaging [1–3], blockchain technology [4], artificial intelligence [5], Internet of Things (IoT) [6], and 5G network [7], intelligent medical system [8] and intelligent diagnosis [9] are becoming more and more popular. However, data security threats [10] make protecting the security of medical data an urgent problem. The volume of medical video data is greater than typical data, which makes the execution of medical data security methods, such as data encryption [11] and integrity detection [12], long. Video retargeting [13] can greatly reduce the storage capacity of video data on the premise of preserving the original content information as much as possible. Medical video retargeting can obtain smaller volume of medical data, then reduce the execution time of data encryption and threat detection algorithm, and improve the performance of medical data security methods.

Traditional image and video retargeting methods, mainly including uniform scaling and direct cropping, only consider the original size and the target size of images, without considering image content. Their effects are unsatisfactory. To improve image and video retargeting performance, researchers proposed content-aware retargeting techniques, which are mainly classified into three types: discrete retargeting [14, 15], continuous retargeting [16–19], and multi-operator retargeting [20–23].

Video retargeting has one more time dimension than image retargeting. It needs to take consideration of the correlation between the contents of adjacent frames. By regarding video as a three-dimensional pixel space-time matrix, Rubinstein et al. proposed FSC [15], looking for and deleting common pixel seams between adjacent frames to eliminate content jitter. NCV [17] proposed by Wolf et al. combines the gradient map, face detection, and foreground motion to produce importance map and then uses mesh deformation to realize video retargeting. Nam et al. [24] proposed a video retargeting method based on Kalman filter and saliency fusion to reduce video content jitter, so as to enhance the robustness of video retargeting. Wang et al. [25] proposed a multi-operator method based on improved seam carving to realize video retargeting. Cho and Kang [26] proposed an interpolation video retargeting method based on image deformation vector network, which uses the displacement vector generated by a convolutional neural network to perform interpolation. Kaur et al. [27] proposed a spatiotemporal seam carving video retargeting method based on Kalman filter.

The existing video retargeting methods mainly focus on the pixel relationship and foreground motion between adjacent frames. These methods aim to ensure the shape of important content in the process of retargeting. However, the above methods do not consider the attention of users to the video content, nor the impact of background movement on retargeting, resulting in serious deformation of the important content or poor quality of retargeting results. Furthermore, the human visual system can quickly find the required information from the visual scene and locate the visual attention to the focus in the scene [28]. Consequently, besides moving objects and important targets, the attention focus also includes the areas where change is about to happen next moment, such as the place where the sun will rise before sunrise, the place where actors will appear on the stage before the performance, and the direction where the ball is moving to.

This paper makes full use of the user's eye tracking data and the motion information of both the background and foreground in the video and proposes a video retargeting method based on visual attention and motion estimation to reduce the deformation of the important area. Firstly, clustering is carried out according to the eye tracking data to generate the visual attention energy map. Then, the motion estimation map is obtained according to the corresponding feature points of the foreground and background between adjacent frames. Thirdly, importance map is generated by composing visual attention energy map, motion estimation energy map, and gradient map. Finally, video retargeting is performed by mesh deformation.

The proposed method utilizes the attention attribute of the human visual system and the movement factor of content in video, so the retargeting result is more in line with people's visual requirements. The experimental results on public datasets show that the method in this paper is better than the compared method in protecting important area and reducing salient object jitter.

2. Proposed Method

As shown in Figure 1, the framework of the proposed VAMEVR (visual attention and motion estimation-based video retargeting) method mainly includes visual attention data clustering, salience detection, SIFT feature detection, motion estimation, mesh deformation, and so on.

2.1. Visual Attention

2.1. Visual Attention. In a video, the areas concerned by the human visual system are usually regarded as important areas. These areas should be of increased energy to reduce deformation in the retargeting process. In this paper, the eye tracking data will be utilized as the basis of visual attention, and it will be abstracted as visual focus. Then, visual attention energy will be generated according to the visual focus.

2.1.1. Visual Attention Focus. This paper takes the eyeball tracking data of DAVSOD [29] dataset as demonstration. As shown in Figure 2, the eyeball tracking data exist in the form of discrete points. Through observation, it is found that most eyeball tracking data points are presented as two clusters.

In this paper, the K-means method [30] is utilized to cluster the eyeball tracking data points into 2 groups. The center of each group is just the visual focus. Firstly, we randomly select 2 data points as the initial cluster centroid. Secondly, we divide the data points into 2 mutually exclusive clusters according to the Euclidean distance from each point to the initial selected data points. Thirdly, the average positions of each cluster are obtained as the new cluster centroid. Then, repeat steps 2 and 3 until the centroid position does not vary.

The example of the focusing result is presented in Figure 3. Figure 3(a) shows the original frame. Figure 3(b) shows the eye tracking data and focusing result. The white point is regarded as the eye tracking data, and the two red points are the centers of two clusters. Figure 3(c) shows the visual attention energy map.

2.1.2. Visual Attention Energy. Visual attention energy indicates the attention of the human visual system to important position in the image. The greater the energy is, the higher the attention is, and vice versa.

Two cluster centroids described in Section 2.1.1 are denoted as $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$. The distances from each pixel of the frame to P_1 and P_2 are separately set as r_1 and r_2. Then, visual attention energy $e(x_i, y_j)$ of each pixel position in the frame is defined as

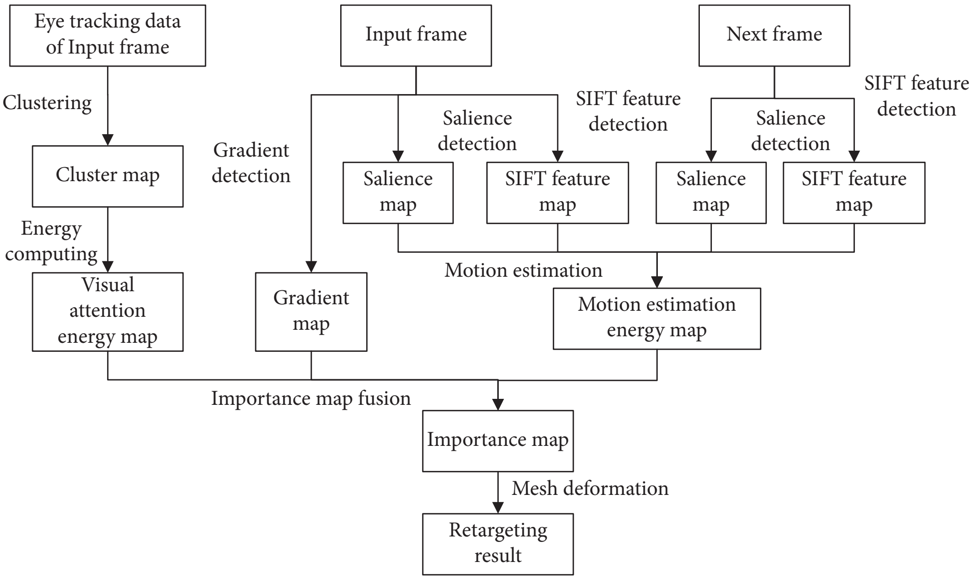


Figure 1: The framework of the proposed VAMEVR method.

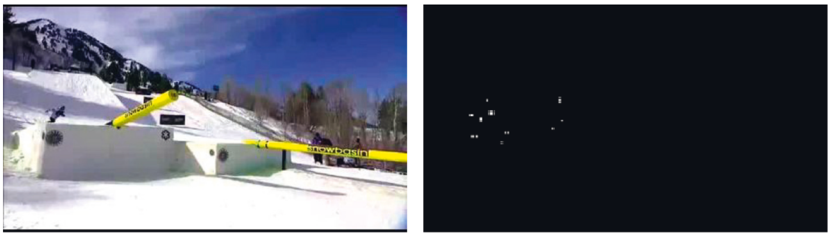


Figure 2: Eye tracking data example in DAVSOD dataset.

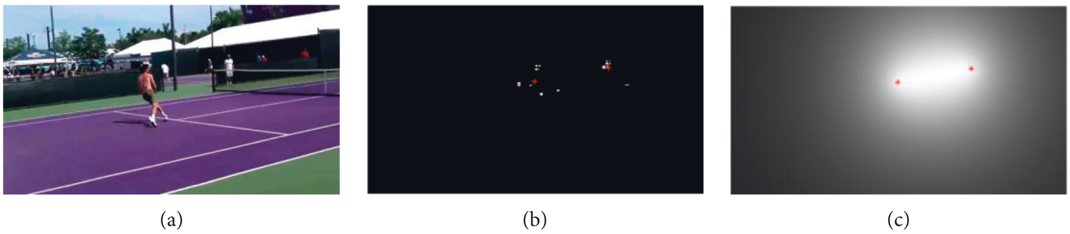


Figure 3: Visual attention energy map.

$$e(x_i, y_i) = \frac{\sqrt{W^2 + H^2}}{r_1 + r_2} = \frac{\sqrt{W^2 + H^2}}{\sqrt{(x_i - x_1)^2 + (y_i - y_1)^2} + \sqrt{(x_i - x_2)^2 + (y_i - y_2)^2}}, \quad (1)$$

where W and H are separately the width and height of the video frame. The generated energy map is shown in Figure 3. Figure 3(c) shows the visual attention energy map, which is generated according to the cluster results of eye tracking data as shown in Figure 3(b).

2.2. Motion Estimation. In a video, the background and foreground are usually moving. In addition, the moving direction and speed of background are different from those of the foreground. The human visual system pays greater attention to the direction where the object is going. For example, in the tennis video, the direction where the players run to will attract more attention. In the racing video, area in front of the car is paid more attention.

Between adjacent video frames, the motion distance and direction of the background and foreground can be calculated to predict the motion trajectory of the salient object. Both current position and the upcoming position of the foreground object are taken as important areas at the same time, which can protect the visual attention areas to reduce the deformation of these important areas in the process of retargeting and improve the visual effect of retargeting results.

2.2.1. Feature Detection. In the background of a video frame, the mean values of displacement of the feature points are used as the base of moving speed. The same is for the foreground of a video. The position to be reached by the foreground significant object is estimated according to the moving speed. Then, both the current position of the foreground and the position to be reached after motion estimation are regarded as important areas.

SIFT (scale-invariant feature transform) [31] is a computer vision algorithm proposed by Lowe to detect regional features in images. The core idea of SIFT algorithm is to find extreme points in multiple spatial scales and calculate position, rotation, light, and scale invariants to describe the features in images. The SIFT algorithm has good robustness, recognition, expansibility, and efficiency.

In this paper, SIFT algorithm feature detection is used to detect the background and foreground motion information between adjacent frames. Also, 20 feature points with the highest reliability are selected as the basis for motion speed calculation. An example of feature points is shown in Figure 4.

2.2.2. Foreground Separation. In a video frame, salient object is generally the foreground area. By salience detection, the foreground area can be separated from the background. Compared with other algorithms, SSAV [29] can obtain clearer and more accurate result. SSAV [29] is mainly composed of a pyramid deconvolution module and salience transfer perception module. The former is used to robustly learn static salience features. The latter combines the traditional long-term memory convolution network with salience transfer perception attention mechanism. This paper uses the SSAV [29] method to separate the salient foreground object from video frames.

2.2.3. Motion Detection and Estimation. From SIFT feature points, we select $n(n = 20)$ point with high reliability as the basis for motion detection and estimation. Concretely, SIFT feature points contained in the background are recorded as $P_{bg}(x_{bg}, y_{bg})$, and the number of those points is n_{bg}. Similarly, SIFT feature points contained in the foreground are recorded as $P_{fg}(x_{fg}, y_{fg})$, and the number is n_{fg}. From frame i to frame $i + 1$, the average moving speed of feature points in the background is recorded as $V_{bg}(dx_{bg}, dy_{bg})$.

$$\begin{cases} dx_{bg} = \dfrac{\sum_{j=1}^{n_{bg}}\left(x_{bg}^{i+1} - x_{bg}^{i}\right)}{n_{bg}}, \\ \\ dy_{bg} = \dfrac{\sum_{j=1}^{n_{bg}}\left(y_{bg}^{i+1} - y_{bg}^{i}\right)}{n_{bg}}. \end{cases} \tag{2}$$

Similarly, from frame i to frame $i + 1$, the average value of the moving speed of the feature points in the foreground is denoted as $V_{fg}(dx_{fg}, dy_{fg})$.

$$\begin{cases} dx_{fg} = \dfrac{\sum_{j=1}^{n_{fg}}\left(x_{fg}^{i+1} - x_{fg}^{i}\right)}{n_{fg}}, \\ \\ dy_{fg} = \dfrac{\sum_{j=1}^{n_{fg}}\left(y_{fg}^{i+1} - y_{fg}^{i}\right)}{n_{fg}}. \end{cases} \tag{3}$$

For a video, the estimated actual motion speed $V_{act}(dx_{act}, dy_{act})$ of the foreground is defined as the difference between the motion speed of the foreground and background.

$$V_{act} = V_{fg} - V_{bg}. \tag{4}$$

Bring equations (2) and (3) into equation (4).

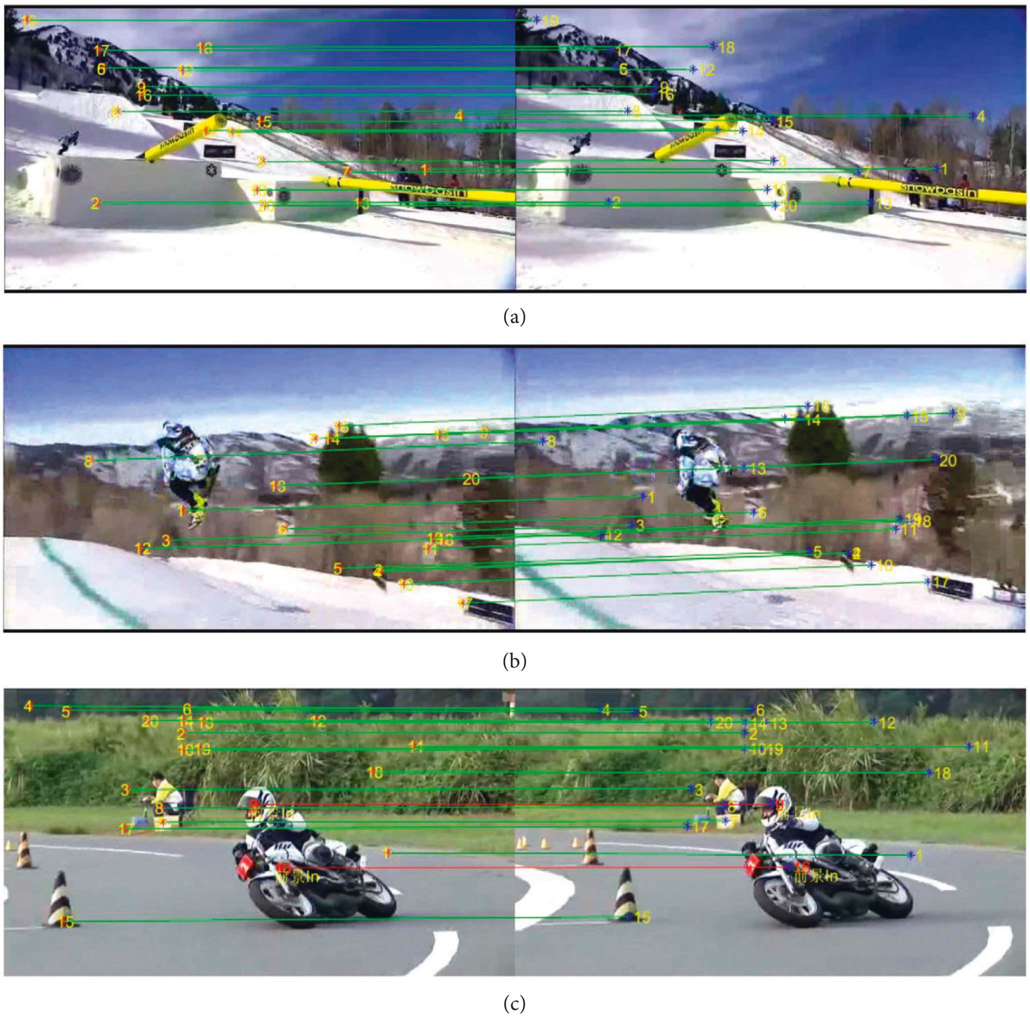
$$\begin{cases} dx_{act} = \dfrac{\sum_{j=1}^{n_{fg}}\left(x_{fg}^{i+1} - x_{fg}^{i}\right)}{n_{fg}} - \dfrac{\sum_{j=1}^{n_{bg}}\left(x_{fg}^{i+1} - x_{bg}^{i}\right)}{n_{bg}}, \\ \\ dy_{act} = \dfrac{\sum_{j=1}^{n_{fg}}\left(y_{fg}^{i+1} - y_{fg}^{i}\right)}{n_{fg}} - \dfrac{\sum_{j=1}^{n_{bg}}\left(y_{bg}^{i+1} - y_{bg}^{i}\right)}{n_{bg}}. \end{cases} \tag{5}$$

As shown in Figure 5, after obtaining the salience map of the current frame, we calculate the edge of the salient region by the Canny [32] method. Then, the edge is overlaid with the actual motion speed $V_{act}(dx_{act}, dy_{act})$ as the predicted position of the salient object. The polygon surrounding method [33] is used to obtain the external polygon of both current and predicted object contour. Finally, the area surrounded by the polygon is just the important region after motion estimation. The motion estimation energy map is the binary map of important area after motion estimation, which is shown in Figure 5(d).

When the salient object is too small or the features are not obvious, the first n $(n = 20)$ feature points detected by the SIFT algorithm are wholly in the background area. In this situation, the centroid displacement of the salience object detected by SSAV is directly used as the moving speed of the foreground object to predict the position where the foreground will go.

The points in salient object area are denoted as $P_{fg}^{*}(xc_{fg}, yc_{fg})$. The number of those points is m_{fg}. From frame i to frame $i + 1$, the motion speed of the foreground's centroid is denoted as $V_{fg}^{*}(dx\ c_{fg}, dy\ c_{fg})$, where

$$\begin{cases} dxc_{fg} = \dfrac{\sum_{l=1}^{m_{fg}^{i+1}}\left(xc_{fg}^{i+1}\right)}{m_{fg}^{i+1}} - \dfrac{\sum_{q=1}^{m_{fg}^{i}}\left(xc_{fg}^{i}\right)}{m_{fg}^{i}}, \\ \\ dyc_{fg} = \dfrac{\sum_{l=1}^{m_{fg}^{i+1}}\left(yc_{fg}^{i+1}\right)}{m_{fg}^{i+1}} - \dfrac{\sum_{q=1}^{m_{fg}^{i}}\left(yc_{fg}^{i}\right)}{m_{fg}^{i}}. \end{cases} \tag{6}$$


(a)

(b)

(c)

Figure 4: SIFT detection between adjacent frames. In (a)–(c), the green lines indicate the connection of feature points in the background, and the red lines indicate the connection of feature points in the foreground. All feature points in (a) and (b) belong to the background area. In (c), only one feature point is in the foreground area and the other feature points are in the background area.

The actual motion speed $V_{act}(dx_{act}, dy_{act})$ of the foreground is the difference between the motion speed of the foreground and the motion speed of the background.

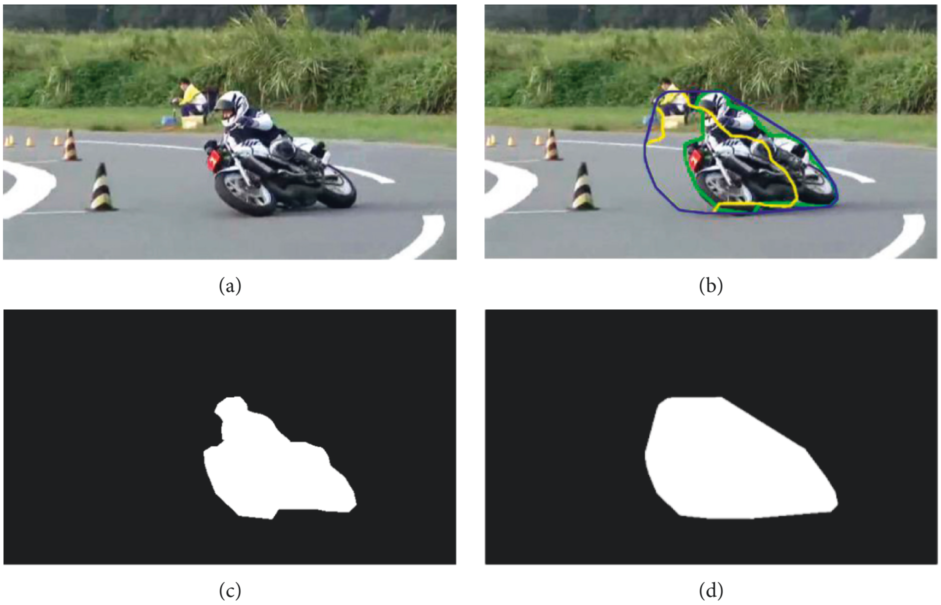
$$V_{act} = V^{*}_{fg} - V_{bg}. \tag{7}$$

Bring (2) and (6) into (7).

$$\begin{cases} dx_{act} = \left(\dfrac{\sum_{l=1}^{m^{i+1}_{fg}} \left(xc^{i+1}_{fg} \right)}{m^{i+1}_{fg}} - \dfrac{\sum_{q=1}^{m^{i}_{fg}} \left(xc^{i}_{fg} \right)}{m^{i}_{fg}} \right) - \dfrac{\sum_{j=1}^{n_{bg}} \left(x^{i+1}_{bg} - x^{i}_{bg} \right)}{n_{bg}}, \\ dx_{act} = \left(\dfrac{\sum_{l=1}^{m^{i+1}_{fg}} \left(xc^{i+1}_{fg} \right)}{m^{i+1}_{fg}} - \dfrac{\sum_{q=1}^{m^{i}_{fg}} \left(xc^{i}_{fg} \right)}{m^{i}_{fg}} \right) - \dfrac{\sum_{j=1}^{n_{bg}} \left(y^{i+1}_{bg} - y^{i}_{bg} \right)}{n_{bg}}. \end{cases} \tag{8}$$

2.3. Importance Map Fusion. The importance map is the direct basis for image retargeting. The visual attention energy map and motion estimation map obtained in the above steps need to be fused into the importance map.

We denote I_{eye} as the normalized visual attention energy map, I_{grad} as the normalized gradient energy map, I_{motion} as the normalized motion estimation energy map, and I_{imp} as the importance map. The coefficient



(a)

(b)

(c)

(d)

Figure 5: Motion estimation. (a) The original frame. In (b), the green contour is the current position of the salient object, the yellow contour is the predicted place of salient object, and the blue contour is the encirclement of both the current and predict areas of salient object after motion estimation. (c) The salience map of (a). (d) The binary map of important area after motion estimation.

$w (0 \leq w \leq 1)$ is the weight of the visual attention energy map in the importance map, over the gradient energy map. Then, the importance map I_{imp} is defined as follows.

$$I_{\text{imp}} = \max\left(\left(w \times I_{\text{eye}} + (1 - w) \times I_{\text{grad}}\right), I_{\text{motion}}\right). \tag{9}$$

The parameter $w (0 \leq w \leq 1)$ determines the visual effect of visual attention energy in importance map. The smaller w is, the smaller the proportion of visual attention energy is. Thus, the impact of visual attention on the results in the retargeting process is smaller, and vice versa. The larger w is, the greater the proportion of visual focus energy is. Therefore, the impact of visual attention on the results in the retargeting process is greater. When $w = 0$, the retargeting results only reflect the gradient information and motion estimation information, not the visual attention information. Also, when $w = 1$, the retargeting results only reflect the visual attention information and motion estimation information, not the visual attention information.

2.4. Mesh Deformation. This paper uses Wang's method [18] for mesh deformation to realize video retargeting. The input frame is divided into quadrilateral mesh (V, E, F). V, E, and F represent the set of vertex, edge, and quadrilateral separately. Each quadrilateral is with a scaling factor s_f. The average importance energy of each quad is w_f. The quad deformation energy is defined as D_u.

$$D_u = \sum_{f\in F} w_f \left(\sum_{(i,j)\,\in\, E(f)} \left\| \left(v'_i - v'_j\right) - s_f\left(v_i - v_j\right) \right\|^2 \right). \tag{10}$$

The grid line bending energy is described as D_l.

$$D_l = \sum_{(i,j\in E)} \left\| \left(v'_i - v'_j\right) - \left(\frac{\left\|v'_i - v'_j\right\|}{\left\|v_i - v_j\right\|} \right) \left(v_i - v_j\right) \right\|^2. \tag{11}$$

The total energy D is the sum of D_u and D_l.

$$D = D_u + D_l. \tag{12}$$

Wang's method [18] uses iterative solver to solve for mesh deformation. In each iteration, the scaling factor s_f of each grid is calculated by local optimization, and then the mesh vertexes are updated by global optimization under the constraint of target image boundary conditions. The iterator will be terminated when the energy is no longer increased or the displacement of mesh vertexes is less than 0.5. The smooth scaling factors s'_f are generated by minimizing the following energy.

$$\sum_{f\in F} w_f \sum_{q\in N(f)} \frac{1}{2}\left(w_f - w_q\right)\left(s'_f - s'_q\right) + \sum_{f\in F} w_f\left(s'_f - s_f\right)^2. \tag{13}$$

2.5. The Algorithm of the Proposed Method. The implementation steps of the proposed methods are shown in Algorithm 1.

3. Results and Analysis

3.1. Experimental Environment and Parameter Settings. To validate the performance of the proposed method, we conduct experiments on a computer with an Intel i7-5500U@2.4 GHz CPU and 16 GB RAM. The proposed method was implemented in MATLAB R2016a on Windows.

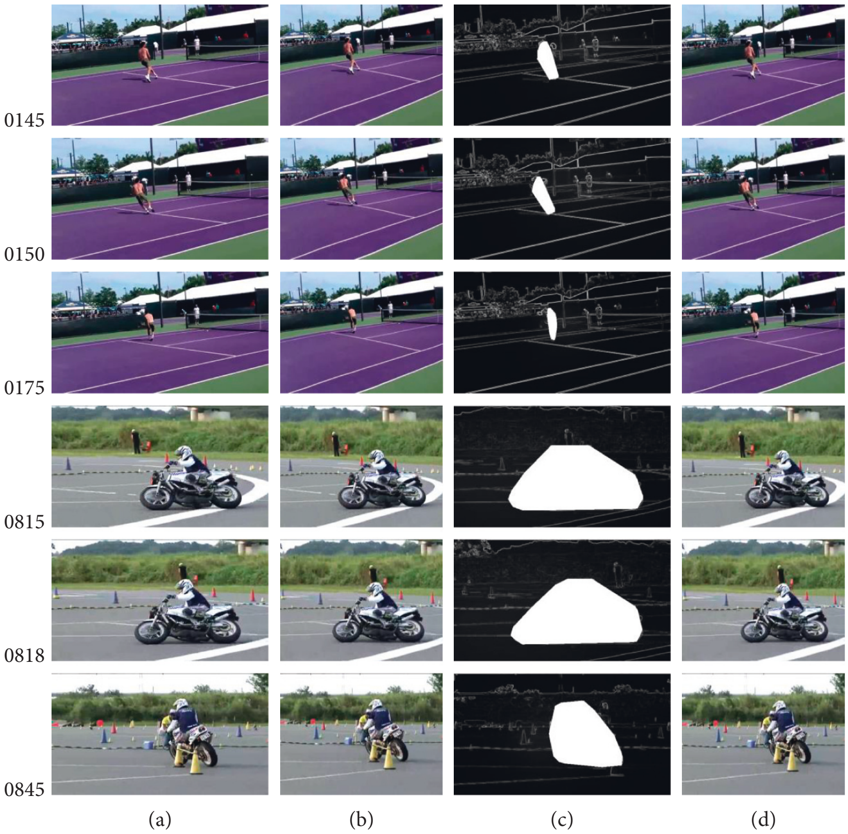


Figure 6: Horizontal retargeting to 75%. (a) Input frame. (b) Result of SNS [18]. (c) Our importance map with $w = 0.1$. (d) Our result with $w = 0.1$.

The number of visual attention data cluster k is set as 2. In the important map fusion process, the weight w of visual attention is set as 0.1, 0.5, and 0.9 separately.

In order to illustrate the universality of proposed method, the public dataset DAVSOD [29] is selected as experimental input. DAVSOD is a large-scale video salient object dataset, which mainly serves the evaluation of video salient object detection and video retargeting. DAVSOD contains 226 video sequences and 24000 frames, covering a variety of scenes, object categories, and motion modes. It is marked strictly according to human eye tracking data.

For each dataset, 3 methods were applied for comparison experiments: forward seam carving (FSC) [15], SNS [18], and the proposed VAMEVR.

3.2. Experimental Result and Analysis. We randomly select "select_0115" and "select_0194" videos of DAVSOD as input data of experiment. The data of "select_0115" include a tennis video clip with 105 frames and 640*360 pixels per frame. The data of "select_0194" include a motorcycle race video clip, with 133 frames in total, and the size of each frame is 640*360. In both of the above video data, the camera is moving during video shooting, that is, the background is moving.

The experimental results are shown in Figures 6 and 7.

From Figures 6 and 7, we can find that the deformation of the salient area is small, especially the area in the direction the object moves to, which is well protected. As shown in Figure 7(d) concretely, the region, where the tennis ball in "0145" video frame is moving to, is with smaller deformation, and so is the area in front of the motorcycle in "0818" video frame.

The main reason of above results is that the important area is of high energy by visual attention and motion estimation. In "0145" and "0150" of Figure 7(c), it can be seen that people paid more attention to the direction of the ball the player was going to move. Similarly, in "0815" and "0818" of Figure 7(c), people pay more attention to the forward direction of the motorcycle and less attention to the rear direction of the motorcycle.

Specifically, as shown in Figures 6(c), 7(a), and 7(c), the smaller w is, the weaker the effect of the visual attention is. The larger w is, the more obvious the effect of visual attention is.

3.3. Time Analysis. The size of video frames and average processing time of each frame in this paper are shown in Table 1.

It can be seen from Table 1 that the time of FSC is longest, with 6.03 s per frame. The average time per frame of VAMEVR in this paper is 0.53 s. It is 0.24 seconds longer than SNS. The increased time is mainly used to calculate visual attention energy and motion estimation detection.

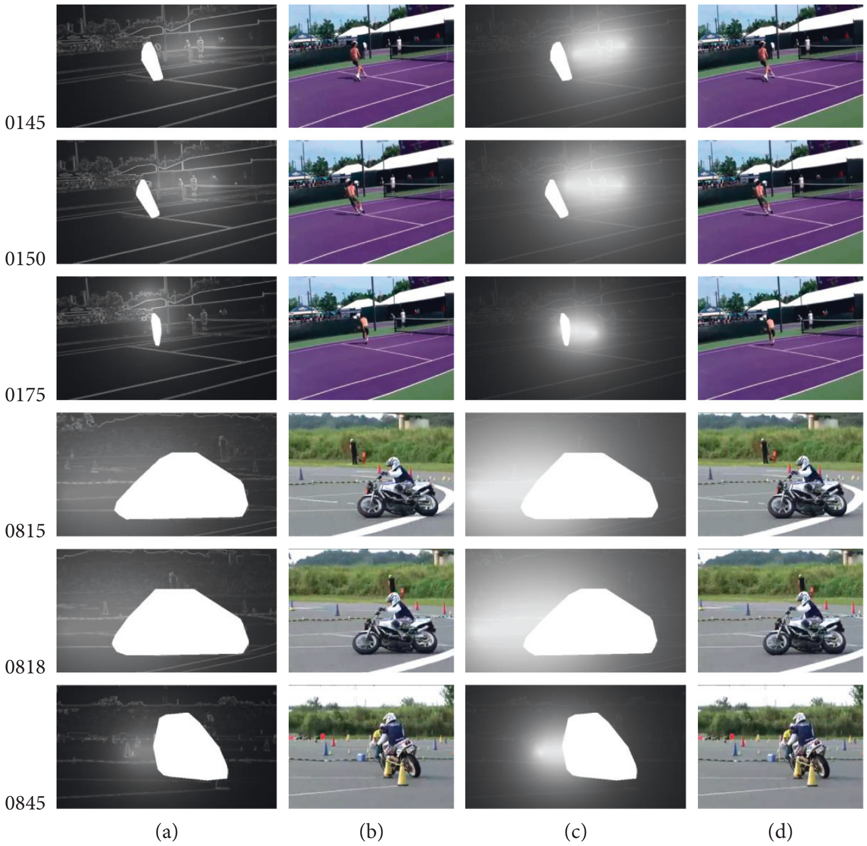


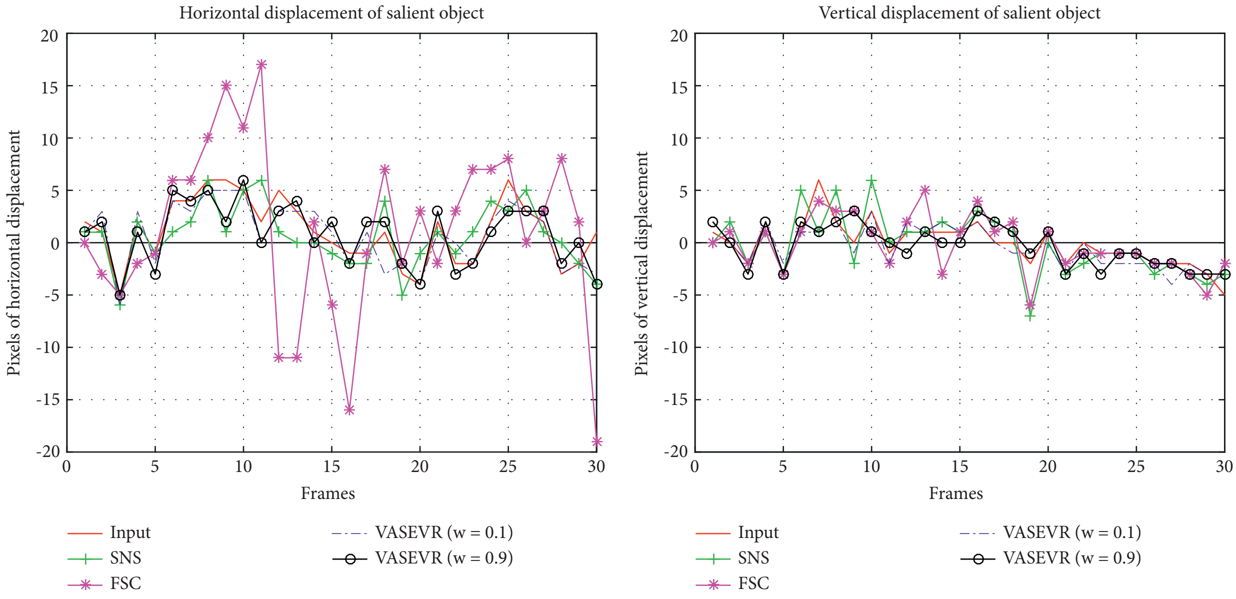
Figure 7: Horizontal retargeting to 75%. (a) Our importance map with $w = 0.5$. (b) Our result with $w = 0.5$. (c) Our importance map with $w = 0.9$. (d) Our result with $w = 0.9$.

```
Input: original video $V_{\text{input}}$, the number of frames $K$, important map fusion coefficient parameter $w$
Output: retargeting result video $V_{\text{result}}$
**For $i = 1$ to $K - 1$ do**
        Calculate the two cluster centers of eye tracking data of $Frame_i$ by K-means method
        Use (1) to produce the visual attention energy map $I_{\text{eye}}$ of $Frame_i$
        Significant object separation of $Frame_i$ and $Frame_{i+1}$ by SSAV [29] model
        Calculate the position of corresponding features of $Frame_i$ and $Frame_{i+1}$ by SIFT method
        Get the number of trusted feature points in foreground and denote it as $n_{fg}$
        If $n_{fg} \geq 1$
            Calculate the background speed $V_{bg}$ between $Frame_i$ and $Frame_{i+1}$ by (2)
            Calculate the foreground speed $V_{fg}$ between $Frame_i$ and $Frame_{i+1}$ by (3)
            Calculate actual moving speed $V_{\text{act}}$ of the salient object by (4) and (5)
        Else
            Calculate the background speed $V_{bg}$ between $Frame_i$ and $Frame_{i+1}$ by (2)
            Calculate the foreground speed $V^{*}_{fg}$ between $Frame_i$ and $Frame_{i+1}$ by (6)
            Calculate actual moving speed $V_{\text{act}}$ of the salient object by (7) and (8)
        End If
        Estimate the position of foreground $(x_{\text{est}\,i}, y_{\text{est}\,i}) = (x_{\text{cur}}, y_{\text{cur}}) + V_{\text{act}}$
        Calculate the circumscribed polygon $R_{fg}$ of both the estimated position and current position of the foreground
        Generate the foreground motion estimation map $I_{\text{motion}}$ according to the salient areas $S_r$ in polygon $R_{fg}$
        Compose importance map $I_{\text{imp}}$ from visual attention energy map $I_{\text{eye}}$, foreground motion estimation map $I_{\text{motion}}$, and gradient map $I_{\text{grad}}$ by (9)
        Use the mesh deformation method described in Section 2.4 to produce retargeting result of $Frame_i$
**End for**
**Output result $V_{\text{result}}$**
```

Algorithm 1: Video retargeting based on visual attention and motion estimation.

Table 1: Execution time of different methods.

Frame size	SNS (s/frame)	FSC (s/frame)	Proposed VAMEVR (s/frame)
640 × 360	0.29	6.03	0.53



(a)

(b)

Figure 8: Comparison of displacement of salient objects before and after retargeting. (a) Horizontal. (b) Vertical.

Table 2: Correlation of salient object displacement before and after retargeting.

Direction methods	SNS	FSC	Proposed VAMEVR		
			$w = 0.1$	$w = 0.5$	$w = 0.9$
Horizontal	0.69185	0.26246	0.89489	0.87823	0.86542
Vertical	0.72326	0.73599	0.84416	0.82361	0.78051

3.4. Discussion. The human visual system is more sensitive to salient objects. The more consistent the displacement of salient objects in adjacent frames before and after retargeting, the lower the content jitter. In this paper, 30 frames of motorcycle racing videos are randomly selected for retargeting.

For the proposed VAMEVR, the centroid displacement of salient object in the retargeting result is basically the same as that of original video. When the weight coefficient of visual attention energy map is 0.1 and 0.9, the comparative analysis of horizontal and vertical displacement is shown in Figure 8.

The displacement correlation of the salient objects can indicate the visual consistency between the original video and the retargeting result. The displacement of the centroid of the significant object in input video and retargeting result is denoted as X and Y separately. The covariance is defined as cov (X, Y), and the standard deviation of X and Y is (σ_X, σ_Y). The Pearson correlation coefficient $\rho_{X,Y}$ is defined as follows.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{14}$$

As shown in Table 2, for VAMEVR, the displacements of the salient objects before and after retargeting are more positively correlated than SNS and FSC. The visual effects of our results are more consistent with the original video than SNS and FSC.

4. Conclusion

This paper proposes a visual attention and motion estimation-based video retargeting method for medical data security. Firstly, clustering is carried out according to the eye tracking data to generate the visual attention energy map. Then, the motion estimation map is obtained according to the corresponding feature points of the foreground and background between adjacent frames. Thirdly, importance map is generated by composing visual attention energy map, motion estimation map, and gradient map. Finally, video retargeting is performed by mesh deformation. Experiments show that the proposed method can protect important area concerned by the human visual system. The displacement of a salient object in retargeting results is more close to input video. Therefore, the visual effect is more in line with human visual need. Our future work is to study the multi-object separation method and then study the video retargeting method based on multi-object motion estimation for medical data security.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Hubei Natural Science Foundation under grant no. 2021CFB156 and the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) under grant no. JP21K17737.

References

[1] F. Ding, G. Zhu, Y. Li, X. Zhang, P. Atrey, and S. Lyu, “Anti-forensics for face swapping videos via adversarial training,” *IEEE Transactions on Multimedia*, vol. 2021, Article ID 3098422, 2021.

[2] A. R. Javed, Z. Jalil, W. Zehra, T. Gadekallu, D. Y. Suh, and M. J. Piran, “A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions,” *Engineering Applications of Artificial Intelligence*, vol. 106, Article ID 104456, 2021.

[3] H. Li, Q. Zheng, J. Zhang, Z. Du, Z. Li, and B. Kang, “Pix2Pix-Based grayscale image coloring method,” *Journal of Computer-Aided Design & Computer Graphics*, vol. 33, no. 6, pp. 929–938, 2021.

[4] W. Wang, Q. Chen, Z. Yin et al., “Blockchain and PUF-based lightweight authentication protocol for wireless medical sensor networks,” *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8883–8891, 2022.

[5] W. Wang, M. H. Fida, Z. Lian et al., “Secure-enhanced federated learning for ai-empowered electric vehicle energy prediction,” *IEEE Consumer Electronics Magazine*, vol. 2021, Article ID 3116917, 2021.

[6] H. Li, Q. Zheng, W. Yan, R. Tao, X. Qi, and Z. Wen, “Image super-resolution reconstruction for secure data transmission in Internet of Things environment,” *Mathematical Biosciences and Engineering*, vol. 18, no. 5, pp. 6652–6671, 2021.

[7] L. Tan, K. Yu, L. Lin et al., “Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space-air-ground integrated intelligent transportation system,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2830–2842, 2022.

[8] A. Balakrishnan, R. Kadiyala, G. Dhiman et al., *A Personalized Eccentric Cyber-Physical System Architecture for Smart Healthcare Security and Communication Networks*, vol. 2021, Article ID 1747077, 2021.

[9] W. L. Shang, J. Chen, H. Bi, Y. Sui, Y. Chen, and H. Yu, “Impacts of COVID-19 pandemic on user behaviors and environmental benefits of bike sharing: a big-data analysis,” *Applied Energy*, vol. 285, Article ID 116429, 2020.

[10] L. Tan, K. Yu, F. Ming, X. Cheng, and G. Srivastava, “Secure and resilient artificial intelligence of Things: a HoneyNet approach for threat detection and situational awareness,” *IEEE Consumer Electronics Magazine*, vol. 11, no. 3, pp. 69–78, Article ID 3081874, 2022.

[11] C. Feng, K. Yu, M. Aloqaily, M. Alazab, Z. Lv, and S. Mumtaz, “Attribute-based encryption with parallel outsourced decryption for edge intelligent IoV,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13784–13795, 2020.

[12] K. Yu, L. Tan, S. Mumtaz et al., “Securing critical infrastructures: deep-Learning-Based threat detection in IIoT,” *IEEE Communications Magazine*, vol. 59, no. 10, pp. 76–82, 2021.

[13] A. Garg, A. Negi, and P. Jindal, “Structure preservation of image using an efficient content-aware image retargeting technique,” *Signal, Image and Video Processing*, vol. 15, no. 1, pp. 185–193, 2021.

[14] S. Avidan and A. Shamir, “Seam carving for content-aware image resizing,” *ACM SIGGRAPH 2007 papers on - SIGGRAPH ‘07*, vol. 26, no. 3, pp. 10–es, 2007.

[15] M. Rubinstein, A. Shamir, and S. Avidan, “Improved seam carving for video retargeting,” *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 1–9, 2008.

[16] Y. Kim, S. Jung, C. Jung, and C. Kim, “A structure-aware axis-aligned grid deformation approach for robust image retargeting,” *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 7717–7739, 2018.

[17] L. Wolf, M. Guttmann, and D. Cohen-Or, “Non-homogeneous content-driven video-retargeting,” in *Proceedings of the IEEE 11th International Conference on Computer Vision*, pp. 1–6, IEEE, Rio de Janeiro, Brazil, December 2007.

[18] Y. S. Wang, C. L. Tai, O. Sorkine, and T. Y. Lee, “Optimized scale-and-stretch for image resizing,” *ACM SIGGRAPH Asia 2008 papers on - SIGGRAPH Asia ‘08*, vol. 27, no. 5, p. 118, 2008.

[19] Z. Karni, D. Freedman, and C. Gotsman, “Energy-based image deformation,” *Computer Graphics Forum*, vol. 28, no. 5, pp. 1257–1268, 2009.

[20] M. Rubinstein, A. Shamir, and S. Avidan, “Multi-operator media retargeting,” *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 1–11, 2009.

[21] A. Garg and A. Negi, “Structure preservation in content-aware image retargeting using multi-operator,” *IET Image Processing*, vol. 14, no. 13, pp. 2965–2975, 2020.

[22] Y. Zhou, Z. Chen, and W. Li, “Weakly supervised reinforced multi-operator image retargeting,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 126–139, 2021.

[23] M. Abhayadev and T. Santha, “Multi-operator content aware image retargeting on natural images,” *Journal of entific and industrial research*, vol. 78, no. 1, pp. 193–198, 2019.

[24] H. Nam, D. Park, and K. Jeon, “Jitter-Robust video retargeting with kalman filter and attention saliency fusion network,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 858–862, IEEE, Abu Dhabi, UAE, September 2020.

[25] S. Wang, Z. Tang, W. Dong, and J. Yao, *Multi-Operator Video Retargeting Method Based on Improved Seam Carving*, IEEE,

in *Proceedings of the IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, pp. 1609–1614, July 2020.

[26] S. I. Cho and S. J. Kang, "Extrapolation-based video retargeting with backward warping using an image-to-warping vector generation network," *IEEE Signal Processing Letters*, vol. 27, no. 1, pp. 446–450, 2020.

[27] H. Kaur, S. Kour, and D. Sen, "Video retargeting through spatio-temporal seam carving using Kalman filter," *IET Image Processing*, vol. 13, no. 11, pp. 1862–1871, 2019.

[28] A. Borji, D. N. Sihite, and L. Itti, "What stands out in a scene? A study of human explicit saliency judgment," *Vision Research*, vol. 91, no. 15, pp. 62–77, 2013.

[29] D. P. Fan, W. Wang, M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8554–8564, IEEE, Long Beach, CA, USA, January 2019.

[30] J. A. Hartigan and M. A. Wong, "Algorithm as 136: a K-means clustering algorithm," *Applied statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[32] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[33] K. Hormann and A. Agathos, "The point in polygon problem for arbitrary polygons," *Computational Geometry*, vol. 20, no. 3, pp. 131–144, 2001.