Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 2, 2023

Research on Video Synopsis Based on Deep Learning Target Detection and Target Trajectory Optimization

Yue Liu[1, a], Li Luo[1, b] and Weibin Hong[2, c]

[1]School of Guangdong Technology University, Guangzhou 510000, China
[2] Guangzhou Guangxin Technology Co. Ltd, Guangzhou 510000, China
[a] 18512513101@163.com, [b] luoli@gdut.edu.cn, [c] hongweibin@guangxin.tech

Abstract: In this era of rapid economic and technological development, monitoring is essential for security issues. However, it is also accompanied by the low efficiency of managing and viewing a large number of surveillance videos every day. Traditional methods are gradually unable to meet the requirements of application, and the video field requires algorithm technology that can better solve practical problems. Intelligent monitoring technology is gradually emerging and developing, based on theories such as computer vision, which can solve many monitoring video problems and greatly improve the efficiency of monitoring video work. Through in-depth exploration of existing video synopsis technology algorithms, this article proposes a new video synopsis method that can effectively detect and track videos, thereby greatly improving the efficiency of storage, transmission, and use of surveillance videos. Experiments have shown that this method can effectively concentrate surveillance videos, and compared with existing methods, it has a better synopsis ratio while ensuring the integrity of video information, effectively reducing collisions between targets, effectively reducing overlap between targets, and achieving good visual effects. Utilizing an improved deep learning object detection and multi object tracking algorithm with added attention mechanism to extract foreground moving targets in videos, and using a mixed Gaussian background modeling algorithm establish a background, laying a stable target foundation for subsequent video concentration. At the same time, design trajectory recombination optimization methods to ensure that the targets do not overlap as much as possible. Reasonably place the targets in the new condensed video sequence, determine the index of all targets in the new condensed video stream, and finally integrate each frame of target images into the background image according to the set index rules, ultimately obtaining the synopsis video.

Keywords: Monitoring video synopsis; Target detection; Deep learning; Track extraction; Trajectory optimization.

1. Introduction

In recent years, with the popularity of the Internet, the powerful functions of computers, and the leap of video data processing technology, the completely digital and networked video monitoring technology has become very important. These technologies have strong openness, integration, and reliable operability, which can effectively improve the quality of video surveillance and promote the progress of the security industry, thus bringing many conveniences to society. With the development of networked video surveillance, intelligent video surveillance has become the mainstream of today's society. It can not only provide more accurate video surveillance, but also achieve high-definition video, providing more reliable guarantees for social management and supervision.

Video condensation technology is a technique that compresses the original video in both time and space, making the length of the condensed video shorter than the original video. In the condensed video, different objects can be combined into a common activity scene. This video concentration technology helps to retrieve specific objects at specific times and locations. This article designs a video concentration method based on object detection and tracking that is suitable for intelligent monitoring systems, ensuring that the ghosting problem in the condensed video is minimized, and that the trajectory of moving objects is reconstructed correctly and easily observed by the human eye, thereby improving work efficiency and reducing losses.

2. Target Detection and Tracking Algorithm Based on YOLOv5 and DeepSortSection Headings

2.1. Modeling of mixed Gaussian background

The use of Gaussian Mixture Model (GMM)[3] modeling technology can effectively analyze the image effect of complex backgrounds and suppress the impact of noise on the target, thereby more accurately fitting the changes in pixel color values. The core idea of mixed Gaussian modeling technology is to use multiple Gaussian distributions and their respective weights to simulate the color changes of different pixels in an image. The advantage of this technology is that it can more accurately capture details in the image, thereby better describing the image. Based on the characteristics of the Gaussian model, multiple Gaussian distributions are arranged according to their corresponding weights, with the highest distribution being considered the optimal background

2.2. YOLOv5 model network structure

The YOLO [4] algorithm refers to a simpler, faster, and more effective one-step output target image fast detection and recognition algorithm, which can directly achieve detection box analysis and extraction of the target to be detected in the input image. Through regression analysis, the algorithm can quickly and accurately detect the target position and category information in the image. For images, we can refine them into several different grids and place them on YOLOv5 [5]'s deep neural network. We will use this model to examine each grid and determine their type based on their characterization. In

addition, we can also use the NMS (Nonlinear Optimization) algorithm to determine the optimal edge conditions.

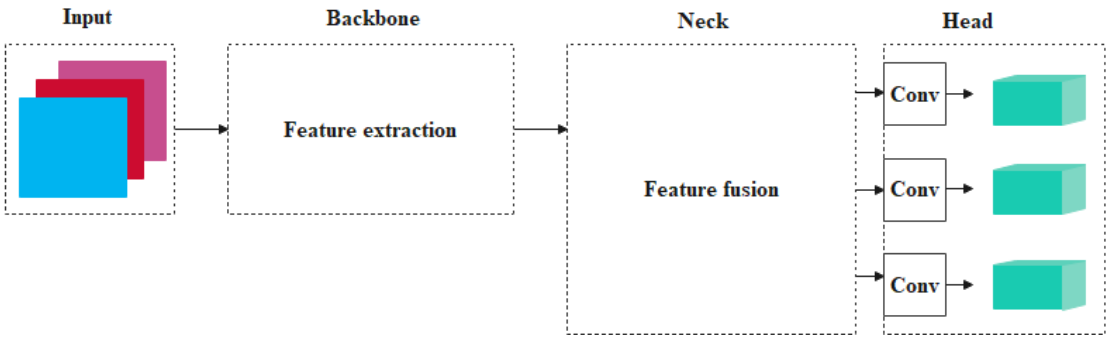

Fig. 1 The Network Structure of YOLOv5

The network structure is mainly divided into four parts: input image, backbone network, neck network and detection head. Mosaic technology can effectively improve the input image, including using techniques such as random scaling, random cropping, and random sorting to effectively combine individual images and significantly improve their quality. This data augmentation method was invented by members of the Yolov5 team and can effectively improve the robustness and generalization ability of the model, as well as increase the number of samples in the training set. By using Mosaic data augmentation technology, not only can images be provided with richer background information, but also the detection accuracy of small targets can be significantly improved. A new automatic anchoring and image adaptation technology has been proposed, which can simultaneously process BN and 4 images, greatly improving the efficiency and accuracy of image segmentation.

2.3. Improvement Based on YOLOv5-Attention Mechanism

The basic principle of Attention is to decompose the source data into several pairs, which are closely related and each has a unique Key and Value. Therefore, in cases where special processing is required on the source data, it is necessary to evaluate these Key, Value and other parameters, and then weight their weight coefficients based on the evaluation results to achieve the expected Attention value. The Attention mechanism aims to utilize the Value value of each element in the Source and obtain the corresponding weight coefficients through precise weighting processing, enabling us to better grasp the importance of each element. According to this formula, we can express its core idea.

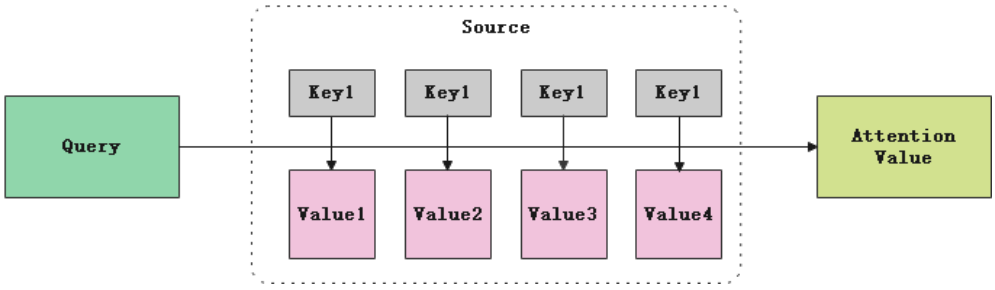

Fig. 2 The encoding process of attention mechanism

$$Attention(Query, Source) = \sum_{i=1}^{L_x} Similarity(Query, Key_i) * Value_i \quad (1)$$

Lx is the length of Source, and the formula meaning is as described above.

2.4. Target tracking algorithm based on DeepSort

DeepSort[6] is a target tracking algorithm based on Kalman filtering and deep learning. The DeepSort algorithm is an improved algorithm that can more finely capture external information of objects, such as optical deformation, noise, vibration, etc. In addition, this algorithm can also achieve fast localization of nearest neighbors and rapid identification of new paths. It is in two states: confirmed or non-confirmed. If the trajectory of an unconfirmed state needs to be continuously matched with the detection box for a period of time or times before it can become a confirmed state. The steps of the DeepSort algorithm include:

1) Generate corresponding trajectories based on the detection results of the first frame. Initialize motion variables using Kalman filtering and predict corresponding boxes. At this point, the trajectory is in an unconfirmed state.

2) Using IOU technology, compare the target box of the previous frame with the trajectory prediction box of this frame, and calculate their cost matrix.

3) After being processed by the Hungarian algorithm, three different results appear: the first is trajectory mismatch, which is deleted after a certain number of times; The second method is to detect the mismatch between the detection box and the prediction box, and update it with a new trajectory; The third method is to ensure that the search frame is completely consistent with the predicted frame, that is, the information in the search frame is correctly captured and the corresponding trajectory is updated after Kalman filtering processing.

4) Repeat steps (2) - (3) until the confirmed trajectory or video frame ends.

5) Through Kalman filtering, we can predict the trajectory boxes of confirmed and non-confirmed states, and use cascade matching technology to determine their relationship. At the same time, we can also use their appearance features, motion information, and pre-reserved 100 frames to achieve this goal.

6) There are three possible outcomes in the matching process of cascading matching:

1. Trajectory matching: Match the detection results in the current frame with the tracking results in the previous frame. If the matching is successful, update the trajectory variables in the Kalman filter.

2. Detection frame mismatch: The detection result in the current frame cannot match the tracking result in the previous frame. At this point, the trajectory of the previously uncertain state is IOU matched with the unmatched detection frame in the current frame to obtain the cost matrix.

3. Trajectory mismatch: The detection result in the current frame cannot match the tracking result in the previous frame. At this point, the previously tracked trajectory is IOU matched with the unmatched detection frame in the current frame to obtain the cost matrix.

By calculating the cost matrix, the optimal matching result can be determined and the variables in the tracker can be updated. The cascade matching method can effectively handle the matching problem in target tracking and improve the accuracy of tracking to a certain extent.

7) Using the Hungarian algorithm, we can obtain three different linear matching results from all the cost matrices obtained: the first is caused by trajectory mismatch, and we can remove these inconsistent trajectories; The second method is due to issues with the detection boxes, which we can reset to new trajectories and use Kalman filtering to perform related transformations; The third method is due to the matching between the detection box and the prediction box, indicating successful tracking of the previous and subsequent frames. We can use Kalman filtering to iteratively update the corresponding trajectory of the corresponding detection box.

8) Repeat steps (5) - (7) until the video frame ends.

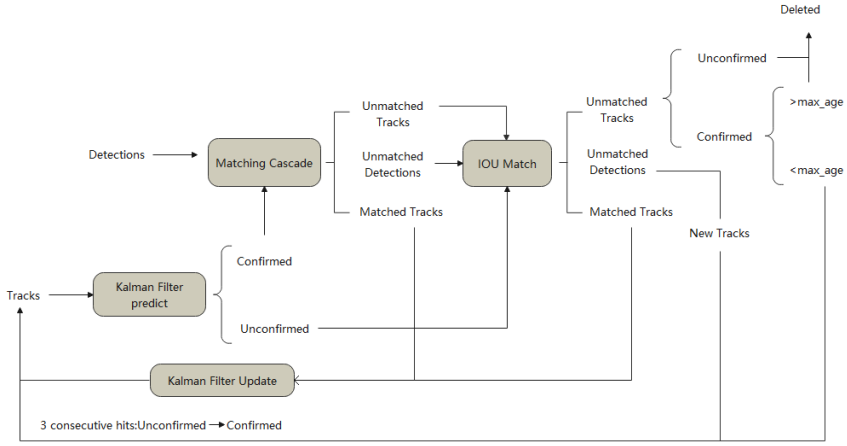

Fig. 3 Flowchart of DeepSort

3. Target trajectory recombination optimization

3.1. Trajectory recombination algorithm

The trajectory reorganization process is to determine the target in the first frame of the new video sequence, traverse all targets to follow a certain logic, and finally reasonably determine the index of the target in the first frame of the new video [7]. Based on the above speed-based batching, the following methods have been set up to minimize target collisions and ultimately determine the time and location of the reconstructed trajectory appearing in the new video stream. The first step is to set the target collision detection sliding window. When adding a new target to a new condensed video sequence, collision detection will be performed on the target within a certain frame range of the previously processed and set target video sequence and the new target to be added to the video sequence that is currently being processed. If there is a collision, no new target will be added to the current frame. Calculate the coordinates of the two-target center of gravity positions from the detection box. When the distance between the horizontal and vertical coordinates of the two-target center of gravity is less than a certain threshold, it will be considered as a collision. The threshold is set to 1.5 to 3 times the average width and average height of the target detection box. The second step is to limit the number of newly added targets in the same frame of the condensed video, and to limit the number of newly added targets in the current frame. After setting the index of the first frame of the condensed video, if no restriction parameters are added, it is highly possible that many targets have the same index of the first frame and a large number of targets appear in the same frame. Finally, it is possible to consider increasing the generation time interval of different batches of targets, and adding new targets to the new condensed video sequence after a slight interval of a certain frame. The condensed video will become slightly longer but the condensed effect will improve. Moreover, the above hyperparameter can be set according to the actual application scenarios, which maximizes the effect of the method in this paper. In summary, the target trajectory has been reorganized in the new video and improved algorithms have been used to achieve better concentration results.

Table 1. Trajectory reorganization algorithm

Trajectory reorganization Algorithm
Input：Target tracking file (targets.json),Target frame position coordinates, Center position, ID, Frame number, Frame length
Output：The target's new index file（new_index.json）
frame_num =0
while account> 0: #Count of all objects that did not find the first frame
for track in target_speed: #Traverse the target files saved in batches with speed respectively
for new_target in new_target_list: # Traverse to find all targets with frame number frame_num
new_target.set_first_frame(frame_num) #Set the first_frame of new_target
video_index[frame_num].add_new_location(new_location) #Then put the position of new_target into new_frame
account -= 1 #Target count minus one
for new_target in new_target_list: # Traverse again the targets that will not have collisions
video_index[frame_num].collision_detect(temp_location) #Set the sliding collision window
new_target.set_first_frame(frame_num)
video_index[frame_num].locations_account < max_target #Limit the number of targets newly added to frame_num
account-=1
frame_num += x #Frame interval x can be set according to the specific scene

The algorithm for the motion target trajectory recombination module is as follows:

(1) Enter the tracking file for the target (targets. json),

including the coordinate information of the target position detection box, the calculation of the target center, the target ID, the frame number where the target appears, and the total frame length where the target appears.

(2) The total loop condition is that the target count (account) for all first frames that have not been found is greater than 0, and all targets are replaced with new_target and save in new_target_list.

(3) First traversal found all frames as frames_num, the target of num is to add new_target , setting the first frame of the target, place the new position in the new frame, and subtract the target count by one.

(4) If the second traversal finds all targets at later times but does not cause conflicts, a sliding window collision detection is performed between the traversal target, the current frame target, and the targets within a certain range of previous frames. If the requirements are met, the secondary target can be added to the current frame, and the target count is reduced by one.

(5) Finally, increase the number of frames x appropriately to reduce the pressure caused by excessive total targets in the current frame.

3.2. Trajectory fusion

Fusion each target image box with the background image, and trajectory fusion is the process of pasting the trajectory with the determined start time into the generated background image. This article uses the Poisson fusion [8] method to fuse images, which seamlessly fuses a portion of the source image onto the target image. Its essence is to generate pixels within the fusion region using the gradient field of the source image as a guide while maintaining the pixels of the target image at the fusion boundary.

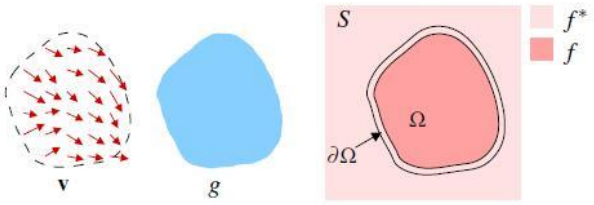

Fig. 4 Flowchart of DeepSort

The overall principle is to keep the gradient field of the generated pixels in the fusion area as consistent as possible with the gradient field of the pixels in the source image fusion part, and to minimize the gradient difference in equation solving. It is to ensure that the Laplace results of the generated region are consistent with those of the source image, and that the values of the boundary of the generated region are consistent with those of the target image in the fusion region. The quadratic optimization result is obtained by discretizing the finite order differential of the Poisson equation.

$$\nabla^2 \varphi = \mathrm{f} \tag{2}$$

$$\min_{f|\Omega} \sum_{|p \cdot q| \cap \Omega} f_q = \sum_{|p \cdot q| \cap \Omega} f_q^* + \sum_{q \in N_p} v_{pq} \tag{3}$$

Among them, p is the pixel point in the source image S, v is the gradient of a certain region in the image, Ω is the target area to be fused placed in the S region, f is an unknown function existing in the Ω region, f* is a known function in S, fq is the value of f at point q, f*q is the value of f* at point q, Np is the four neighboring regions of p,<p, q>is a pixel pair, vpq is the mapping result of v [(p+q)/2] on [p, q], and its solution satisfies the formula.

$$\left|N_p\right| f_p - \sum_{q \in N_{p \cap \Omega}} f_q = \sum_{q \in N_r \cap \partial\Omega} f_q^* + \sum_{q \in N_p} v_{pq} \tag{4}$$

fp is the value of f at point p. The steps for Poisson fusion in condensed video are as follows:

(1) Calculate the gradient field of the external rectangular box of the moving target and the gradient field of the background image.

(2) Calculate the gradient field of the image after the fusion of the moving target and the background image, and then directly cover the gradient field of the moving target onto the gradient field of the background image, while obtaining the gradient value of each pixel.

(3) Calculate the divergence of the fused image by taking the partial derivative of the pixel gradient value obtained in step (2).

(4) Solve the coefficient matrix of the Poisson reconstruction equation to obtain the final fusion result.

4. Summary

4.1. Experimental Environment Design

In this study, we evaluated the feasibility of using object detection ghosting optimization for video concentration technology under surveillance conditions through two sets of controlled experiments. These two sets of experiments were both conducted using Pycharm software and completed in Python 3.7 and OpenCV4.5.2.54 programming, and within the learning framework of Python. The system adopts Intel (R) Core (TM) i5-4210U CPU and has a running speed of 2.7GHz. Memory 8GB, Windows 10 operating system.

4.2. Experimental results and analysis

To verify the performance of the concentration method, this chapter uses five segments of pedestrian surveillance videos as input videos. Video 1 is a commonly used video dataset for object detection, with pedestrians being relatively dense. The rest are self-made video datasets, which are videos of pedestrians walking in different scenes. In videos 2 and 4, there are fewer and smaller foreground targets, while in videos 3 and 5, there are slightly more and larger foreground targets. The following figure shows the original screenshot, background modeling, and detection tracking information. The figure shows a condensed video of three methods from video one to video five.

In order to quantitatively evaluate the concentration effect more objectively and accurately, this article uses the following three concentration performance indicators: (1) compactness rate (space utilization rate) CR [9] is expressed as the ratio of all target pixel regions in each frame of the concentrated video to all pixel regions in each frame of the concentrated video. (2) COR (Compact ratio) is an important parameter that reflects the degree of video compression, which is the ratio of the number of frames in the original video to the number of frames in the condensed video. When the concentration ratio is high, it can effectively improve the compression effect, thereby maximizing the utilization of space and time. (3) The degree of "pseudo collision" shows that due to the rearrangement of target trajectories, targets that could not have collided were condensed into the video, resulting in a false correlation level (FCL) [10]. The degree of 'pseudo collision' is even more pronounced, as it can be used to measure the mutual influence between two objects

and their correlation.

$$\mathrm{fcl}(n)=\frac{\sum_{i=1}^{M}\sum_{j=i+1}^{M}tsc(b_i,b_j,n)}{\sum_{i=1}^{M}\chi_{b_i}(x,y,t)} \tag{5}$$

$$\mathrm{tsc}(b_i,b_j,n)=\begin{cases}\sum(\chi_{b_i}(x,y,n-\hat{t}_i^s+t_i^s)\cdot\chi_{b_j}(x,y,n-\hat{t}_i^s+t_i^s)) & n\in\hat{t}_j\cap\hat{t}_i\\ 0 & \text{else}\end{cases} \tag{6}$$

$tsc(b_i,b_j,n)$ is the size of the occluded area between two targets and the nth frame in the condensed video, M is the total number of frames, and the denominator of the formula is the sum of all foreground target pixels in the i-th frame. When the collision area is larger or the target area in the frame is smaller, the numerator of the above formula is larger, and vice versa. At the same time, the video synopsis method (DL)[9] integrating deep learning target recognition and the video synopsis method (TC) [12] based on fill density and dynamic programming are given, and compared with the experimental results of the concentration method (GO) based on target detection ghost optimization proposed in this paper.

Fig. 5 Experimental results of this method

Fig. 6 Experimental results of this DL

Fig. 7 Experimental results of this TC

Table 2. Five video experiment results

Video 1	COR	FCL	CR	Original frame count	Number of frames after synopsis
GO	9.80	0.13	0.32	8863	904
DL	11.98	0.35	0.27	8863	740
TC	15.30	0.43	0.21	8863	579
Video 2					
GO	2.92	0.01	0.09	1124	384
DL	2.99	0.03	0.08	1124	375
TC	3.25	0.04	0.06	1124	346
Video 3					
GO	3.56	0.01	0.25	1975	554
DL	4.26	0.09	0.20	1975	463
TC	4.89	0.25	0.09	1975	403
Video 4					
GO	4.57	0.01	0.16	1510	330
DL	4.66	0.01	0.17	1510	325
TC	5.07	0.02	0.14	1510	298
Video 5					
GO	1.57	0.05	0.21	755	480
DL	2.11	0.19	0.11	755	358
TC	2.40	0.25	0.09	755	315

According to the calculation and analysis of Video 2 and Table Video 4, there are fewer collision rates in scenes with fewer targets. From Video 1, Video 3, and Video 5, it can be concluded that when the scene is denser than the target and the target is larger, the compared method has more collisions, resulting in a higher compactness rate, which is not as effective as the method in this article. The collision rate of our method is lower. From the results, it can be concluded that

although the concentration ratio is not optimal, it is more uniform in time without losing some video concentration time, greatly reducing the collision rate between moving targets, effectively reducing the phenomenon of "pseudo collision", making every effort to ensure that the targets do not overlap and have better visual effects, which is more applicable to a wide range of scenes than other methods.

Acknowledgements

At this point in writing, there are countless emotions. With the passage of time, the two and a half years of graduate studies are coming to an end. In these three years of study and life, Guangdong University of Technology has given me too many beautiful memories and gained many sincere friendships. The motto of the school, "Morality, ambition, good learning and good deeds", will become the Guiding Light of my future life. Looking back, all the stories from those who first entered the laboratory to now are vivid, with too much reluctance and unforgettable lingering in my heart. As I am about to enter society, I am filled with emotions and grateful to the teachers, classmates, and family who have accompanied me all the way.

The National Natural Science Foundation of China (11574058).

References

[1] Jian Rong Cao,Yang Xu,Cai Yun Liu. Algorithm of Surveillance Video Synopsis Based on Objects[J]. Applied Mechanics and Materials,2013,2388(321-324).

[2] Michael G. Christel,Michael A. Smith,C. Roy Taylor,David B. Winkler. Evolving video skims into useful multimedia abstractions[P]. Human Factors in Computing Systems,1998.

[3] HUANG X,ZHOU J,LIU B. Moving targets detection approach based on adaptive mixture Gaussian background node[J].Journal of Computer Applications,2010,30(1):128-131.

[4] Joseph Redmon,Santosh Kumar Divvala,Ross B. Girshick,Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection.[J]. CoRR,2015,abs/1506.02640.

[5] Luo Yu,Zhang Yifan,Sun Xize,Dai Hengwei,Chen Xiaohui. Intelligent Solutions in Chest Abnormality Detection Based on YOLOv5 and ResNet50.[J]. Journal of healthcare engineering,2021,2021.

[6] Jie Yang,Leonidas LilianAsimwe,Mumtaz Farhan,Ali Munsif. Ship Detection and Tracking in Inland Waterways Using Improved YOLOv3 and Deep SORT[J]. Symmetry,2021,13(2).

[7] Thirumalaiah G.,Pandian S. Immanuel Alex. An Optimized Novel Technique for Video Synopsis Using Bayesian Object Tracking Algorithm[J]. Journal of Computational and Theoretical Nanoscience,2020,17(11).

[8] P erez, M. Gangnet, and A. Blake. Poisson image editing. In ACM Transactions on graphics (TOG), volume 22, pages 313–318. ACM, 2003.

[9] Tian Cai,Zhe Lin. A Method for synopsis Surveillance Video Abstracts by Integrating Deep Learning Target Recognition [J]. Modern computers,2020(24):49-53.

[10] Lin Zhang. Research on Video Synopsis Technology Based on Adaptive Tracking Evaluation Mechanism [D]. Shandong University,2017.

[11] Chao Cong. Surveillance video summarization algorithm based on dynamic programming and fill density [J]. Computer Engineering,2018,44(07):250-258.DOI:10.19678/j.issn.1000-3428.0047716.