Video Synopsis Based on Attention Mechanism and Local Transparent Processing

SHENGBO CHEN, XIANRUI LIU, YIYONG HUANG, CONGCONG ZHOU, AND HUAIKOU MIAO, (Member, IEEE)

1School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
2Shanghai Key Laboratory of Computer Software Testing and Evaluating, Shanghai 201112, China

Corresponding author: Shengbo Chen (schen@shu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61572306 and Grant 61502294, in part by the National Key Research and Development Program of China under Grant 2017YFB0701600, in part by the High Performance Computing Center, Shanghai University, and in part by the Shanghai Engineering Research Center of Intelligent Computing System under Grant 19DZ2252600.

ABSTRACT The increased number of video cameras makes an explosive growth in the amount of captured video, especially the increase of millions of surveillance cameras that operate 24 hours a day. Since video browsing and retrieval is time consuming, while video synopsis is one of the most effective ways for browsing and indexing such video that enables the review of hours of video in just minutes. How to generate the video synopsis and preserve the essential activities in the original video is still a costly and labor-intensive and time-intensive work. This paper proposes an approach to generating video synopsis with complete foreground and clearer trajectory of moving objects. Firstly, the one-stage CNN-based object detecting has been employed in object extraction and classification. Then, combining integrating the attention-RetinaNet with Local Transparency-Handling Collision (LTHC) algorithm is given out which results in the trajectory combination optimization and makes the trajectory of the moving object more clearly. Finally, the experiments show that the useful video information is fully retained in the result video, the detection accuracy is improved by 4.87\% and the compression ratio reaches 4.94, but the reduction of detection time is not obvious.

INDEX TERMS Video synopsis, attention mechanism, transparency processing, deep learning.

I. INTRODUCTION

Video synopsis is a task with important research value and practical significance. With the development of $7 \times 24$ video surveillance system, the amount of surveillance video data is growing sharply. How to conduct video browsing and retrieval efficiently becomes a challenging work. Especially, in road monitoring or pedestrian monitoring, some moving objects needs to be identified and analyzed in a very long video. The traditional method is to browse the video by manually controlling the playback, which costs too much time. In addition, some important information will be missed due to the people’s own reasons. Original videos contain a large number of “static” frames, each frame only contains a background, without any moving objects. In a sense, static frame is short of useful information, such as object moving trajectory. So it is necessary to design a method to shorten the browsing time of a large number of “static” pictures while completely retaining the useful information in the video. Hence the technology of video synopsis was born.

The earliest video synopsis technologies, such as storyboard [1], scene graph [2], and video manga [3], similar to the video “skimming” or video fast forward, are realized by selecting the “key frames” which contains more information in the video, then recombining the key frames into the synopsis video in chronological order. The main drawback of these existing methods is that they can’t keep the complete trajectory of the moving object in the video, so the result of the synopsis video is not conducive to browse. With the advance of video processing technology, object-based video synopsis [4] has emerged, which focus on the motion of objects in the video. It uses “activity” as the processing unit, while the “key frame”-based video synopsis uses “key frame”. Activities can be shifted on the temporal domain, and
Video synopsis is the simultaneous presentation of events that enables the review of hours of video footage in just minutes, that is, how to show all the moving objects in the original video in a short time. According to the different forms of the result of video synopsis, video synopsis can be divided into the key-frame-based video synopsis and object-based synopsis. This paper focuses on object-based video synopsis, which is also known as dynamic video synopsis.

Object-based video synopsis is an activity-based video condensation technique. The aim of the object-based video synopsis is to display as many activities as possible simultaneously in the shortest time period [5]. The main feature of this technique is that activities from different periods can be shift into the same frame by analyzing the object motion. Thus an efficient condensation performance is achieved.

Object-based video synopsis was proposed by Rav-Acha et al. [13], Pritch et al. [4], [10], [14]. The method consists of on-line part and off-line part. The former generates activities and the latter rearranges the activities. Their study is important because they provide a general framework of video synopsis for the first time. Most of the subsequent research work focuses on the steps presented in Fig.1. In terms of object detection and tracking, Pritch et al. [13] use pixel difference with temporal median method, which is a basic algorithm to exact foreground. Some studies such as [15]–[18] et al. apply the background modeling method
based on statistics like GMM to the object detection of video synopsis. Besides, Mahapatra et al. [19] contribute to the field by using human detection method because this method provide more precise results as the false detection ratio is lower. It’s worth noting that Jin et al. [20] use R-CNN with Chi-square distance for the video synopsis, where CNN-based object detection method has been used in video synopsis. CNN-based detector makes the object extracted more complete in video synopsis. And inspired by their work, this paper uses one stage object detection instead of two stage method, which results in faster speed and better effect for video synopsis. Optimization method is the key point for activity rearrangement and it aims to find the best which aims to find best arrangement of the activities in order to display most of them and avoid collision as far as possible. Energy function is designed and minimized by optimization algorithm. Simulated annealing algorithm [21] is the most commonly used algorithm, which is used in many studies [4], [14], [22]–[25]. After that, mean-shift [15], genetic algorithms [26] and the packing cost [13] based optimization methods are used. These off-line optimization methods are very slow in the condition of too many activities. On-line methods rearrange the activities while detecting them, and this kind of optimization method requires less memory and reduces computational complexity. An online step wise optimization has been done in endless video [27], which adopts an online content-aware approach in a step-wise manner, and sticky tracking is used to achieve high-quality visualization with less computational cost. Huang et al. [28] use table-driven approach, they design a synopsis table to represent each pixel in a video frame with index of detected object and frame. The table would be updated while the objects or the frames are detected. The shortcoming of on-line optimization method is that it cannot achieve a higher compression ratio than that of the off-line method.

In recent years, with more and more researchers joining in the study, the research on video synopsis task has made new progress. Ahmed et al. [29] give a method to generate video synopsis by grouping tracklets of moving objects using spatio-temporal relation which is meaningful. After that, they come up with another method by adding three layers CNN to classify the objects in video to generate synopsis according to a query given by users, and the classified video synopsis is realized in their paper [30]. Some studies focus on object rearrangement and collision handle: Nie et al. [31] handle the collisions by hanging the speed or size of the objects, and develop a Metropolis sampling algorithm to find the solution for three-variable optimization problem. Ruan et al. [32] propose a dynamic graph coloring algorithm to efficiently rearrange all tubes by determining when they should appear, and they use online optimization in the method. In this paper, we also provide a method which realizes the classify video synopsis, the difference from the work in [30] and that in this paper is that we use the labels generated by CNN-based detector itself to classify objects instead of adding an extra network to do classification work, and it reduces the amount of additional neural network design work and computation of the algorithm.

B. CNN-BASED OBJECT DETECTION AND OBJECT SEGMENTATION

Object Detection and Object Segmentation are two main tasks of computer vision. Both tasks can be used to identify areas of interest from images or videos. The early object detection methods distinguish “foreground” of interest and “background” of no interest by generating a mask, which is not much different from the current object segmentation task. However, with the development of deep learning, the term “object detection” tends to describe the technology that uses bounding box to identify and classify targets, while “object segmentation” implies to classify each pixel ins an image as a foreground object.

Traditional object detection methods, such as HOG [33], SIFT [7] and DPM [34] detector, usually use hand-crafted features to detect objects. With the rise of convolutional neural networks(CNN), CNN-based object detection method has become the most mainstream research direction in this field. CNN is a useful filter which can learn robust and multi-levels features of the input photo. CNN-based object detection method can be divided into two directions: two-stage detection and one-stage detection. The former uses anchor calibration and object classification, while the later uses the method completely in on step [9]. Regions with CNN features network(R-CNN) [35] is the most representative algorithm of two-stage methods. It produces object candidate boxes by using selective search [36] at first, then extracts features of boxes by CNN and classification. Under the guidance of this idea, R-CNN develops a variety of improved versions, among which the best one is Faster-RCNN [37], whose main contribution is Region Proposal Network(RPN). RPN uses a CNN to replace selective search. Faster R-CNN is the first end-to-end deep learning detector (VOC2007 mAP = 73.2%, VOC2012 mAP = 70.4%), and it breaks through the speed bottleneck of R-CNN. However, two-stage detectors are computation redundancy, and it leads to the low detection speed.

Different from those two-stage detectors, one-stage CNN detection methods detect an image at once, so they have faster speed. YOLO [38] is a typical one-stage method. It uses one network both to divide proposal region and calculates the probability of belonging to a certain class. The speed of detection is rapidly improved due to the reduction of computation. However, the increase in speed comes at the expense of accuracy. Single Shot MultiBox Detector (SSD) [39] introduces the multi-reference and multi-resolution detection techniques to improve the accuracy of one-stage detectors. A loss function of object detection called “focal loss” and a RetinaNet network have been given out which solve the imbalance of the classes in the task [40]. Focal loss and RetinaNet make the one-stage detectors achieve the same ability as two-stage detectors in accuracy and keep the merit of high speed while detecting as a one-stage method. Therefore, this framework
is a suitable component to perform video synopsis task. Section III introduces the theory of RetinaNet in detail, and we refer to the backbone of RetinaNet and merge it with attention mechanism in order to achieve a higher accuracy.

Semantic segmentation is a computer vision task with a higher semantic level than the “object detection”. It gives the classification of objects as each pixel in the image. Specifically, the input image is converted into a mask with highlighted areas of interest. When using CNNs for semantic segmentation, the encoder-decoder structure is commonly used, and the most classical structure is Fully Convolutional Network (FCN) proposed by Long et al. [41]. It uses deconvolution to upsampling the feature. Besides, DeepLab [42] puts forward atrous convolution, and the authors add Conditional Random Field (CRF) at the end of the network. Related techniques of semantic segmentation are also used in more refined object detection, such as Wang et al. [43] use FCN and proposed deep video saliency network which consists two modules to capture the spatial and temporal saliency information in order to realize video saliency training and detection.

On the basis of semantic segmentation, video segmentation has rapidly developed. In recent years, more and more studies have been done on video segmentation. According to the difference of supervision required, the conventional video segmentation techniques can categorize into unsupervised, semi-supervised and supervised methods. Among them, unsupervised and semi-supervised methods are the hot research directions. These two types of methods can split the object of the video clip with less guidance information. The semi-supervised method uses the object segmentation information of the first frame, while the unsupervised method has no prior information. Caelles et al. [44] propose OSVOS model, which realized semi-supervised video target segmentation by training CNN serval times on different datasets and combining with pixel-wise sigmoid balanced cross entropy loss function. The MaskTrack model is developed by Perazzi et al. [45]. It is also a semi-supervised method by combining off-line and on-line training strategies and using temporal information of the video to guide training. Wang et al. [46] come up with “super-trajectory” to represent a group of compact point trajectories which has similar appearances and close spatiotemporal relationships to guide to propagate the initial annotations in the first frame onto the remaining frames accurately. Ventura et al. [47] use ConvLSTM module and design a encoder-decoder module, incorporating recurrence on spatial and temporal domains to realize a video object segmentation network for semi-supervised and unsupervised tasks. Song et al. [48] design pyramid dilated convolution module and pyramid dilated bidirectional ConvLSTM, and they cascade them to realize video salient object detection task.

III. PROPOSED METHOD

In this section, we provide all components and methods we used in this paper.

A. THE METHOD FOR IMPROVING OBJECT DETECTION

ACCURACY OF VIDEO ABSTRACT

Extracting the moving object in the video is the first step in video synopsis. When using surveillance camera, the video will inevitably have jitter and illumination change [49]. Tradition methods cannot exact object clearly because the fuzzy and exposure in the video impact the result of the traditional object foreground detectors. Fig. 2(b) shows that tradition methods cannot identify moving objects or shadow changes. Currently, CNNs are the most researched algorithms in image analysis. When the feature pictures are input, CNNs can preserve spatial relationships in the pictures by filtering operation. CNN can detect objects completely through training on the specific datasets and not affected by illumination change or shadow. Fig. 2(c) shows the result of object detection using CNN-based method.

It is noted that in this paper we use bounding box based object detection method instead of object segmentation method. We do this after considering the following points: First, video in a single scene has little change of video background, so when detecting the object with bounding box based object detection method, the residual background pixels in boxes lead to little noise for the entire synopsis video. Second, based on the current research, bounding-box based object detection method has a higher accuracy than object segmentation method, and the use of object segmentation method will lead to the incomplete object exaction, which will affect the quality of synopsis video. Third, from the perspective of model training, the production of the custom dataset of bounding box object detection method is easier than that of the object segmentation method, and the computation of the former is smaller than that of the latter.

Based on the above considerations, we use a RetinaNet-based object detection method to exact foreground. RetinaNet is a one-stage CNN object detection algorithm which use “Focal Loss” to improve the imbalance of foreground-background during training. Although two-stage approaches like R-CNN obtain higher accuracy, but the speed of detection is low because they use anchor selection to generate candidate region, and a lot of time is spent when generate synopsis video. RetinaNet is able to match the speed of previous one-stage detectors while surpassing the accuracy.
of state-of-the-art two-stage detectors like Faster R-CNN. The focal loss is defined as:
\[
\text{FL}(p_t) = -\alpha (1 - p_t)^\beta \log(p_t)
\]
In (1), \(p_t\) is defined as:
\[
p_t = \begin{cases} p & \text{if } y = 1, \\ 1 - p & \text{otherwise.} \end{cases}
\]
where \(\alpha, \beta\) are hyperparameters in the model. Focal loss solve the imbalance of positive/negative class and easy/hard examples by modifying the traditional cross entropy loss function. The authors of focal loss argue that class imbalance is the primary reason caused lower accuracy of one-stage object detection methods than two-stage method, and focal loss makes up for this shortcoming [40]. The accuracy of object detection is improved while the speed is guaranteed.

To further improve the accuracy and avoid the condition of object miss, we incorporate the visual attention mechanism into RetinaNet and rename it as “attention-RetinaNet”. Visual attention mechanism is a method to imitate human visual attention, it guides the feature extraction model to focus the valuable part of the picture and ignore the invaluable part. Visual attention mechanism can be used for a variety of computer vision tasks, such as object detection [50] and video object segmentation [51]. Two main aspects of attention mechanism are proposed for feature extraction: channel attention mechanism and spatial attention mechanism. The nature of using attention mechanisms to improve object detection accuracy is that attention mechanism enables the CNNs to focus on the more informative features of the image. In this paper, we use two different attention modules to transform the raw RetinaNet structure: CBAM module and SE-Layer.

Convolutional Block Attention Module(CBAM) [52] exploits both channel attention mechanism and spatial attention mechanism to improve the performance of feature extraction, and has two sub-modules named “channel attention module” and “spatial attention module”. Fig. 3 shows the whole structure and the schema of each module in CBAM. CBAM can be summarized as following formulas:
\[
\begin{align*}
F_{ca} &= CA(F_{in}) \otimes F_{in} \\
F_{out} &= SA(F_{ca}) \otimes F_{ca}
\end{align*}
\]
where \(F_{in} \in \mathbb{R}^{H \times W \times C}\) is the feature picture produced by feature extract network such as ResNet [53] or Inception [54]. CA(·) represents channel attention module and SA(·) is spatial attention module. Channel attention module in CBAM emphasizes the importance of max pooling and the authors combine it with average pooling in the CA(·) function. CA(·) function is given by the following formula:
\[
CA(F) = \sigma \left( A_2 \delta \left( A_1 \frac{F_{avg}}{F_{max}} \right) + A_2 \delta \left( A_1 F_{max} \right) \right)
\]
where \(\text{AvgPool}(\cdot)\) presents average pooling, \(\text{MaxPool}(\cdot)\) means max pooling, \(\delta\) refers to ReLU function, \(\sigma\) refers to Sigmoid function, \(A_1, A_2 \in \mathbb{R}^{C \times C}, r\) is the reduction ratio. Then, a SA(·) function is operated on the output of CA(·) function:
\[
\text{SA}(F) = \sigma \left( f_7^7 \cdot \left( \text{AvgPool}(F); \text{MaxPool}(F) \right) \right)
\]
where \(f_7^7\) represents a convolution operation with the filter size of \(7 \times 7\). CBAM learns what and where to emphasize or suppress, and refines intermediate features effectively [52].

Squeeze-and-Excitation(SE)-Layer was proposed in the study of [55]. Contrary to CBAM module, it only emphasizes the importance of channel attention mechanism. The schema of the SE-Layer is shown in Fig. 4. It can be seen that this structure is similar to the channel attention module of CBAM module, but in the SE-Layer, only average pooling is used.
instead of maxpooling. The author of CBAM module argue that maxpooling gathers the information about distinctive object features while average pooling aggregating spatial information [52]. Since the structure is similar, the mathematical expression of SE-Layer will not be repeated here.

Both SE-Layer and CBAM module are lightweight modules, that means they do not cause too much computational cost. These two ideas inspire us to make a happy marriage between RetinaNet and attention modules. Hence, we design an RetinaNet-based network to perform the object detection task in video synopsis. The overall framework of our approach is shown in Fig. 5.

The framework has four main components. First, we separately add two kinds of attention modules mentioned above to modify residual module in ResNet which is responsible for computing a convolutional feature map over the transformed image. ResNet produces features at different scales \((c_3, c_4, c_5, p_6, p_7)\). The first three feature graphs \((c_3, c_4, c_5)\) are passed through FPN [56], then the obtained feature graphs and the last two feature graphs generated by ResNet are combined to build the feature pyramid \((p_3, p_4, p_5, p_6, p_7)\). The reason for using FPN is that multi-scale features can be used to detect objects of different scales, thus obtaining higher accuracy [40]. Afterwards, we use box subnet and class subnet given in [40] to predict anchor boxes and classes of objects. The box subnet has four \(3 \times 3\) convolution layers and ReLU layers, followed by a \(3 \times 3\) convolution layer with \((class \times anchor)\) channels. The class subnet is similar, but the last convolution layer is \((4 \times anchor)\) channels instead. In Section IV, we compared the effects of two attention modules on the performance of the above framework through experiments, give the experiments process and the analysis of the results.

\[
\text{FIGURE 5. The structure of attention-RetinaNet.}
\]

B. TRAJECTORY OPTIMIZATION FOR VIDEO SYNOPSIS

In stage of activity rearrangement, the balance between video compression and moving object collisions should be considered. Since multiple objects in different video frames are moved in same frame, moving object collisions is inevitable. These collisions will lead to chaos in the scene of synopsis video. Usually, a clustering-based track combination optimization method is used to solve the problem. The method defines an energy function to represent conflict and lost in spatial and temporal domain [5]. Optimization algorithms such as simulated annealing algorithm are used to optimize the energy function for spatiotemporal position planning of moving objects. Some other algorithm like “traffic light” algorithm [57] let one object stop moving and wait for another object to avoid collision in synopsis video, but it can lead to track interruption which can make people misunderstand that whether the object is moving or static. Table 1 gives an overview of existing research of collision handling for video synopsis, we group the methods with similar idea, then we compare characters of each group.

Inspired by the idea of Semi-Transparency-Handling Collision (STHC) method [49], we use a transparent strategy to handle collisions, and propose Local Transparency-Handling Collision (LTHC) algorithm. The researchers of [49] use semi-transparent to handle collision, but when two moving objects collide, they suddenly change from visible to translucent. In particular, when a very small object or a misdetected “object” (which may actually be the background) collides with another large real object, it can create a large area of transparency in the moving object, and it might affect the visual perception for person. To that end, we improve the equation of STHC as follows:

\[
P_{\text{out}}(x,y) = \begin{cases} 
    p_{\text{in}1}(x,y) & (x,y) \in \text{in}1 \setminus \text{in}2 \\
    \gamma p_{\text{in}1}(x,y) + (1-\gamma)p_{\text{in}2}(x,y) & (x,y) \in \text{in}1 \cap \text{in}2 \\
    p_{\text{in}2}(x,y) & (x,y) \in \text{in}2 \setminus \text{in}1 
\end{cases}
\]

where \text{in}1 and \text{in}2 are two objects exacted by detector, \(p_{\text{in}1}(x,y)\) is the pixel value at position, \text{out} is the result of our method.
TABLE 1. Existing research of collision handling for video synopsis.

| Methods Types          | Existing Studies                                      | Overview of Methods | Characteristic                                                                 |
|------------------------|-------------------------------------------------------|---------------------|--------------------------------------------------------------------------------|
| Optimization-based     | Offline methods: simulated annealing [21], genetic algorithms [26], and graph-based methods [58] et al. | Design a collision cost function as a part of activity energy function and find optimal solution for activity collision. | Function design is challenging, and the video synopsis maybe longer than source video if the spatial collision is completely prevented especially for the crowded scenes. |
| Activity-based methods | Traffic light algorithm [57]                         | Let one object stop moving and wait for another object to avoid collision in synopsis video. | Sometimes lead to track interruption which can let people misunderstand that whether the object is moving or static. |
| Transparency-based      | STHC [49]                                            | Two moving objects are treated with global semi-transparency when collision happen. | If there is more noise in the object detection, it will lead to too many translucent areas, affecting the quality of synopsis video. |

Algorithm 1 gives the detailed description of our LTHC method.

In Fig. 6, we provide a schematic diagram of different transparent strategy to handle collisions. (a) is collision with no transparent. (b) uses the global transparency in [49], and (c) uses the our LTHC algorithm. It shows that there is severe occlusion in (a), which makes it impossible to display the object (the red car in the image). (b) the occlusion part can be shown, however, the rest of the part is not clear due to the use of translucent. (c) let the occlusion part be transparent, while the other part keeps the original pixel value, so as to solve the collision occlusion while not affecting the clarity of other parts.

Algorithm 1  Local Transparency-Handling Collision Algorithm

**Input:** Exacted objects sets \( T = \{ T_1, T_2, \ldots, T_n \} \).

**Output:** Synopsis video

1. for \( i = 0 \) to \( n \) do
2. for all \( p_{in_1} \) in \( T_i \) do
3. for all \( p_{in_2} \) in \( j \)th frame do
4. Calculate \( p_{out} \) using 7;
5. Stick \( p_{out} \) into \( j \)th frame;
6. Write \( j \)th frame into output video file;
7. Set \( p_{out} \leftarrow p_{in_2} \) in \( j \)th frame
8. end for
9. end for
10. end for

IV. EXPERIMENTS

We use the optimization methods of video synopsis proposed in this paper. First we present the experimental results of attention-RetinaNet. Second we analysis the result of video synopsis with LTHC. What’s more, we use attention-RetinaNet and LTHC to realize a whole classified video synopsis system.

A. EXPERIMENTS OF ATTENTION-RetinaNet

In order to train our model and compare the performance of the object detectors, we do a series of comprehensive experiments on different datasets. We perform experiments on a server with two Intel(R) Xeon(R) E5-2650 v3 2.3GHz CPUs and an NVIDIA Tesla M60 GPU with an 8GB memory. All of our experiments are implemented in Python with PyTorch.

We use the pretrained ResNet-50 as the backbone of the RetinaNet and adopt the following parameters in the RetinaNet-based models: In (1), \( \alpha = 0.25, \beta = 2 \). In (5), the reduction ratio \( r = 16 \). In the training, batch size is 6, total epoch number is 100, and use SGD optimizer with learning rate is 0.001 with the momentum of 0.9 and weight decay with \( L_2 \) regularization set the parameter of \( 5 \times 10^{-4} \). We re-scaled the images and resized the short side to 600. In NMS module, we set the threshold as 0.45.

In the first experiment, in order to compare performance of four detection models as YOLO v3, RetinaNet, RetinaNet with CBAM, and RetinaNet with SE-Layer, we performed on the public datasets, PASCAL VOC 2007. The result is shown in Table 2. We calculated the mean average precision (mAP) and average precision (AP) of each class to help us understanding the impact of attention module. The attention module enhances the feature contrast of CNNs, and the improved feature can get effective result of accuracy of object detection. Comparing the two attention modules, because CBAM uses a combination of spatial and channel attention mechanism, it shows a greater ability to improve the accuracy of object detection than that of SE-Layer.

Further, we design a dataset within about 5,000 pictures, which because of the next video synopsis experiment.
The data come from the video clips in a traffic road video data set [60], which contains serval videos from traffic surveillance cameras. Video frames are extracted from these video data and the object detection data are marked. We tagged the marked objects into three classes: “ca”, “truck”, and “bus”. The data is divided into training set and verification set by a ratio of 7:3. The number of object of each class in the training set are shown in the Table 3. Meanwhile, these objects have different levels of size. During detection, objects of different sizes can be detected and classified.

Table 4 shows the results of different object detection network with our designed dataset. In this part we compare three models: RetinaNet, RetinaNet with CBAM, and RetinaNet with SE-Layer. The first row represents the result of raw RetinaNet on the dataset. The second row shows the result of the RetinaNet with CBAM module. The last row shows the result of RetinaNet with SE-Layer. We can get similar conclusion as we get in the first experiment.

Table 5 shows the speed of each method when detect on videos with the size of 1920 × 1080. It could be seen that although the attention mechanism increased the computational complexity of the model, the introduction of the two modules did not reduce the speed too much. Even with the
addition of the CBAM module, the detection time of each frame was still more than 3 fps, and the accuracy improved by 4.87%.

The visualization of the object detection shows in Fig. 7. Column (a): the input images, (b): the detection result of raw RetinaNet, (c): the detection result of RetinaNet with CBAM module, (d): the detection result of RetinaNet with SE-Layer. In this figure, we choose three typical objects to show: the first row represents incomplete objects in the frame. The second row is a lot of objects of different sizes. The last row shows the big object in one frame. Through our experiment, we found that RetinaNet with CBAM module has the following two advantages: First, for incomplete objects, this structure reflects better detection results (see the first row of Fig. 7). Second, for smaller targets, the network can “notice” and detect them, thanks to the use of the attention mechanism (see the second row of Fig. 7).

### B. EXPERIMENTS OF LTHC METHOD

For trajectory optimization, LTHC algorithm can improve the compression ratio. While using LTHC method, two different objects can be seen in the same position at the same time, and the overlap does not block each other. We conduct our method in (7) with \( \gamma = 0.5 \). Fig. 8 shows a collision between two objects: two cars happen collision in the synopsis video. The two vehicles from different frames in the original video (so they occur at different time), the collision not really happens in the original video but happens when generate synopsis video. Column (a) is the result of our LTHC method and column (b) is the result of STHC. We find that our method effectively solves the problem of unclear object in synopsis video.

Without the use of complex optimization algorithms, the compression ratio of videos is also greatly increased. We calculate the compression ratio \( (CR) \) by the following equation:

\[
CR = \frac{t_{\text{original}}}{t_{\text{output}}} 
\]

where \( t_{\text{original}} \) refers to the duration of the original video and \( t_{\text{output}} \) is the duration of the output synopsis video. The result of the compression ratio is shown in Table 6. Also, we compare the compression ratio using method in the study of [59] and the result is in Table 7. It can be seen that our method can get a higher compression ratio.

---

**TABLE 3. Number of per classes.**

| Class Name | Number |
|------------|--------|
| Car        | 4652   |
| Truck      | 1021   |
| Bus        | 262    |

**TABLE 4. Accuracy of each detection method using our dataset.**

| Method Name   | mAP  | AP of Car | AP of Truck | AP of Bus |
|---------------|------|-----------|-------------|-----------|
| RetinaNet     | 60.37| 81.9      | 80.2        | 19.0      |
| RetinaNet + CBAM | 63.26| 84.1      | 80.6        | 25.1      |
| RetinaNet + SE-Layer | 62.63| 83.1      | 78.6        | 26.2      |

**TABLE 5. Average detect time and speed of each detection method.**

| Method Name   | Average Detect Time per Frame(s) | Speed(fps) |
|---------------|----------------------------------|------------|
| RetinaNet     | 256.76                           | 3.89       |
| RetinaNet + CBAM | 308                            | 3.24       |
| RetinaNet + SE-Layer | 257.94                        | 3.24       |
C. THE IMPROVEMENT EFFECT OF THE METHODS PROPOSED

Since CNN-based object classifiers detect the objects at the same time, it is feasible to realize classified video synopsis in our method. We compare the effects of previous video synopsis method with ours, and the effects show in Fig. 9. Since CNN-based object detectors classify the objects at the same time, it is feasible to realize classified video synopsis. We compare the effects of previous video synopsis method with ours, and the effects show in Fig. 9. In our method, we improve the quality of video synopsis in object detection and activity rearrangement respectively. CNN-based one-stage object detection algorithm has been applied to the field of video synopsis. Also, STHC algorithm is improved as LTHC algorithm has been applied to the field of video synopsis. (e)(f) shows ours. We have improved in the following three aspects: First, the use of RetinaNet-based object detection method keeps the integrity of detected objects. The synopsis video generated by the traditional object detection algorithm will be incomplete and the light changes will be detected as moving objects by mistake, resulting in the low quality of the synopsis video (show in (c)). However, our method can greatly reduce the impact of this problem. Second, the labels of objects can help us generate video synopsis of specific types of objects. (e)(f) are shown the result when car or truck is specified. Third, the LTHC method let the object more clearly in the synopsis video than that of STHC method.

V. CONCLUSION AND FUTURE WORK

In this paper, in order to get a better synopsis video, we proposed two methods. First, attention-RetinaNet object method, Second, LTHC method. These two methods improve the quality of video synopsis in object detection and activity rearrangement respectively. CNN-based one-stage object detection algorithm has been applied to the field of video synopsis. Also, STHC algorithm is improved as LTHC algorithm, and the unclear object problem is improved under the premise of ensuring the high compression ratio of the original algorithm.

In the future, we will analyze how to raise the speed of detector based on CNN to accelerate the video synopsis task. Also, the parallel computing and big data technology can be considered for this task to reduce the time spending on the video synopsis task and the cost of storage space in the computing system.

ACKNOWLEDGMENT

The authors thank the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

REFERENCES

[1] M. G. Christel, A. G. Hauptmann, A. S. Warmack, and S. A. Crosby, “Adjustable filmstrips and skims as abstractions for a digital video library,” in Proc. IEEE Forum Res. Technol. Adv. Digit. Libraries, May 1999, pp. 98–104.
[2] B.-L. Yeo and B. Liu, “Rapid scene analysis on compressed video,” IEEE Trans. Circuits Syst. Video Technol., vol. 5, no. 6, pp. 533–544, Dec. 1995.
[3] S. Uchihashi, J. Foote, A. Girogsonih, and J. Boreczky, “Video manga: Generating semantically meaningful video summaries,” in Proc. 7th ACM Int. Conf. Multimedia. New York, NY, USA: Association for Computing Machinery, 1999, pp. 383–392, doi: 10.1145/319463.319654.
[4] A. Rav-Acha, Y. Pritch, and S. Peleg, “Making a long video short: Dynamic video synopsis,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, Jun. 2006, pp. 435–441.
[5] K. B. Baskurt and R. Samet, “Video synopsis: A survey,” Comput. Vis. Image Understand., vol. 181, pp. 26–38, Apr. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314219300190
[6] C. Harris and M. Stephens, “A combined corner and edge detector,” in Proc. 4th Alvey Vis. Conf., Jan. 1988, pp. 147–151.
[7] D. G. Lowe, “Distinctive image features from scale-invariant key-points,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004. doi: 10.1023/B:VISI.0000029664.99615.94.
[8] C. P. Papageorgiou, M. Oren, and T. Poggio, “A general framework for object detection,” in Proc. 6th Int. Conf. Comput. Comput. Vis., Jan. 1998, pp. 555–562.
[9] Z. Zhu, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: A survey,” 2019, arXiv:1905.05055. [Online]. Available: http://arxiv.org/abs/1905.05055
[10] Y. Pritch, A. Rav-Acha, and S. Peleg, “Nonchronological video synopsis and indexing,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 11, pp. 1971–1984, Nov. 2008.
[11] H. Li, Y. Huang, and Z. Zhang, “An improved faster R-CNN for same object retrieval,” IEEE Access, vol. 5, pp. 13665–13676, 2017.
[12] Q. Zhang, Y. N. Wu, and S.-C. Zhu, “Interpretable convolutional neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8827–8836.
[13] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg, “Clustered synopsis of surveillance video,” in Proc. 6th IEEE Int. Conf. Adv. Video Signal Based Survell. Sep. 2009, pp. 195–200.
[14] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg, “Webcam synopsis: Peeking around the world,” in Proc. IEEE 11th Int. Conf. Comput. Vis., Oct. 2007, pp. 1–8.
[15] M. Xu, S. Z. Li, B. Li, X.-T. Yuan, and S.-M. Xiang, “A set theoretical method for video synopsis,” in Proc. 1st ACM Int. Conf. Multimedia Inf. Retr. MIR. New York, NY, USA: Association for Computing Machinery, 2008, pp. 366–370, doi: 10.1145/1460096.1460156.
[16] M. Li, Y. Wang, and G. Pan, “Generating fluent tubes in video synopsis,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2013, pp. 2292–2296.
[17] C.-H. Hsia, J.-S. Chiang, and C.-F. Hsieh, “Low-complexity range tree for video synopsis system,” Multimedia Tools Appl., vol. 75, no. 16, pp. 9885–9902, Aug. 2016, doi: 10.1007/s11042-015-2714-2.
[18] X. Li, Z. Wang, and X. Lu, “Surveillance video synopsis via scaling down objects,” IEEE Trans. Image Process., vol. 25, no. 2, pp. 740–755, Feb. 2016.
[19] A. Mahapatra, P. K. Sa, B. Majhi, and S. Padhy, “MVS: A multi-view video synopsis framework,” Signal Process., Image Commun., vol. 42, pp. 31–44, Mar. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596516000059
[20] J. Lin, F. Liu, Z. Gan, and Z. Cui, “Online video synopsis method through simple tube projection strategy,” in Proc. 8th Int. Conf. Wireless Commun. Signal Process. (WCSP), Oct. 2016, pp. 1–5.
[21] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” Science, vol. 220, no. 4598, pp. 671–680, 1983.
[22] M. Rodriguez, “CGRAM: Compact representation of actions in movies,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 3328–3335.
S.-Z. Wang, Z.-Y. Wang, and R.-M. Hu, “Surveillance video synopsis in the compressed domain for fast video browsing,” *J. Vis. Commun. Image Represent.*, vol. 24, no. 8, pp. 1431–1442, Nov. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320313001818

R. Zhong, R. Hu, Z. Wang, and S. Wang, “Fast synopsis for moving objects using compressed video,” *IEEE Signal Process. Lett.*, vol. 21, no. 7, pp. 834–838, Jul. 2014.

W. Liao, Z. Tu, S. Wang, Y. Li, R. Zhong, and H. Zhong, “Compressed-domain video synopsis via 3D graph cut and blank frame deletion,” in *Proc. Thematic Workshops ACM Multimedia-Thematic Workshops*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 253–261. doi: 10.1145/3126808.3126707

Y. Tian, H. Zheng, Q. Chen, D. Wang, and R. Lin, “Surveillance video synopsis generation method via keeping important relationship among objects,” *IET Comput.*, vol. 10, no. 8, pp. 868–872, Dec. 2016.

S. Feng, Z. Lei, D. Yi, and S. Z. Li, “Online content-aware video condensation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*. Washington, DC, USA: IEEE Computer Society, Jun. 2012, pp. 2082–2087.

C.-R. Huang, P.-C.-J. Chung, D.-K. Yang, H.-C. Chen, and G.-J. Huang, “Maximum a posteriori probability estimation for online surveillance video synopsis,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1417–1429, Aug. 2014.

A. Ahmed, S. Kar, D. P. Dogra, R. Patnaik, S. Lee, H. Choi, and I. Kim, “Video synopsis generation using spatio-temporal groups,” in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Sep. 2017, pp. 512–517.

S. Ahmed, D. P. Dogra, S. Kar, R. Patnaik, S. C. Lee, H. Choi, G. P. Nam, and I.-J. Kim, “Query-based video synopsis for intelligent traffic monitoring applications,” *IEEE Trans. Intell. Transp. Syst.*, early access, Jul. 26, 2019, doi: 10.1109/TITS.2019.2929618.

Y. Nie, Z. Li, Z. Zhang, Q. Zhang, T. Ma, and H. Sun, “Collision-free video synopsis incorporating object speed and size changes,” *IEEE Trans. Image Process.*, vol. 29, pp. 1465–1478, 2020.

T. Ruan, S. Wei, J. Li, and Y. Zhao, “Rearranging online tubes for streaming video synopsis: A dynamic graph coloring approach,” *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3873–3884, Aug. 2019.

N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Soc. Comput. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1–8.

J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, “Selective search for object recognition,” in *Proc. 21st Int. Conf. Pattern Recognit.*, Aug. 2012, pp. 1956–1959.

S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017. doi: 10.1109/TPAMI.2016.2577031.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 10, no. 2, pp. 154–171, Sep. 2017. doi: 10.1145/2945116.2939782.

T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

W. Wang, J. Shen, and L. Shao, “Video salient object detection via fully convolutional networks,” *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.

S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixe, D. Cremers, and L. Van Gool, “One-shot video object segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 221–230.
XIANRUI LIU is currently pursuing the master’s degree with the School of Computer Engineering and Science, Shanghai University (SHU). His current research interests include video synopsis, video understanding, big data, and cloud computing.

YIYONG HUANG is currently pursuing the master’s degree with the School of Computer Engineering and Science, Shanghai University (SHU). His current research interests include action recognition, video captioning, and video synopsis.

CONGCONG ZHOU is currently pursuing the master’s degree with the School of Computer Engineering and Science, Shanghai University (SHU). His current research interests include image processing and deep learning.

HUAIKOU MIAO (Member, IEEE) is currently a Professor of computer engineering and science with Shanghai University (SHU), China. His research interests include formal methods and software engineering.