applied sciences

MDPI

Article

Local Attention Sequence Model for Video Object Detection

Zhenhui Li [1], Xiaoping Zhuang [1], Haibo Wang [2,3], Yong Nie [4,*] and Jianzhong Tang [4]

[1] College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China; lizhenhui@zju.edu.cn (Z.L.); 21832070@zju.edu.cn (X.Z.)
[2] Technology and Equipment of Rail Transit Operation and Maintenance Key Laboratory of Sichuan Province, Chengdu 610031, China; haibowang@home.swjtu.edu.cn
[3] School of Mechanical Engineering, Southwest Jiaotong University, Chengdu 610031, China
[4] The State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University, Hangzhou 310027, China; jztang@zju.edu.cn
* Correspondence: ynie@zju.edu.cn

Abstract: Video object detection still faces several difficulties and challenges. For example, the imbalance of positive and negative samples leads to low information processing efficiency, and detection performance declines in abnormal situations in video. This paper examines video object detection based on local attention to address such challenges. We propose a local attention sequence model and optimized the parameter and calculation of ConvGRU. It could process spatial and temporal information in videos more efficiently and ultimately improve detection performance under abnormal conditions. The experiments on ImageNet VID show that our method could improve the detection accuracy by 5.3%, and the visualization results show that the method is adaptive to different abnormal conditions, thereby improving the reliability of video object detection.

Keywords: video object detection; local attention; sequence model; ConvGRU

Citation: Li, Z.; Zhuang, X.; Wang, H.; Nie, Y.; Tang, J. Local Attention Sequence Model for Video Object Detection. *Appl. Sci.* **2021**, *11*, 4561. https://doi.org/10.3390/app11104561

Academic Editor: Sungho Kim

Received: 26 March 2021
Accepted: 4 May 2021
Published: 17 May 2021

1. Introduction

Object detection is a fundamental problem in computer vision and has been widely used in the fields of surveillance, robots, medical intelligence, etc. In recent years, with the rapid popularization of deep convolutional networks, many scholars have also conducted much research on object detection algorithms based on deep convolutional networks [1,2], which has greatly improved the performance of image object detection. In 2013, Girshick proposed R-CNN [3], which used a deep convolutional network to achieve object detection for the first time, and the mAP index on the VOC dataset was approximately doubled when compared to traditional detection methods. Subsequently, Ren proposed Faster R-CNN [4] and designed a region proposal network to extract candidate regions, which improves the detection accuracy and greatly reduces the detection running time. However, the speed of the two-stage detection model based on R-CNN is low. Redmon proposed the YOLO [5] detection framework in 2015, which rasterizes images and predicts the object category and bounding box for each grid at the same time.

Applying such image-based object detectors to the domain of videos, however, is often unsatisfactory due to the deteriorated appearance caused by issues such as motion blur, out-of-focus camera, and rare poses frequently encountered in videos. These problems cannot be effectively solved by relying only on static images. The video can provide context and temporal information containing multiple frames of images. Combining this information can solve the above problems more effectively. Existing methods that leverage temporal information for object detection from videos usually use optical flow to propagate high-level features across frames. Extra optical flow models, e.g., FlowNet [6,7], have to be utilized to establish motion information and achieve better performance, which leads to excessive model parameters and calculations that are not conducive to model deployment.

In addition, the optical flow models establish motion information between local pixels, and it is difficult to model the continuity between high-level semantic features.

Instead of relying on optical flow, we propose an innovative video object detection model based on local attention. Specifically, we design the spatial attention module and local attention sequence model to improve video object detection accuracy and modify ConvGRU (convolutional gated recurrent units) [8] to process video context and temporal information in order to improve object detection reliability.

We conducted extensive experiments on ImageNet VID for video object detection. Our results outperform the original method in accuracy and achieve real-time detection in the same time duration.

In summary, our contributions are as follows:

We introduce a novel video object detector based on local attention to establish the spatial and temporal correspondence across frames without extra optical flow models.

We propose a spatial attention module and local attention sequence module and modify ConvGRU to model spatial and temporal appearance and enhance feature representation.

We conduct experiments on ImageNet VID and achieve improved performance.

2. Related Work

2.1. Image Object Detection

Existing state-of-the-art methods for image object detection mostly follow two paradigms, that is, two-stage and single-stage pipelines. A two-stage pipeline consists of region proposals, region classification, and location refinement. Girshick proposed the R-CNN [3] detection framework, which extracts image features through a convolutional network. Subsequently, Ren proposed Faster R-CNN [4], designed an RPN network based on a convolutional network, and introduced multi-scale anchor boxes to extract candidate regions with higher confidence. Lin proposed FPN [9] to detect objects on a multi-stage feature map. Comparing to two-stage detectors, single-stage methods are faster but less accurate. Redmon proposed YOLO [5], predicting the categories and bounding boxes on each grid simultaneously. Liu proposed SSD [10], which uses anchors on the feature maps of different depths of deep convolutional networks and then obtains categories and bounding boxes through convolution operations on each layer of the maps. Lin proposed focal loss [11] function to address the imbalance of easy and hard examples. In this paper, we use YOLO as our base detector.

2.2. Video Object Detection

The T-CNN [12] framework designed the multi-context suppression module and motion-guided propagation module to process context and motion information between adjacent frames and combine tracking algorithms in order to improve the classification accuracy of detection sequences. The Seq-NMS [13] algorithm only incorporates video temporal information in the post-processing operation of image object detection and can significantly improve the performance of video object detection through simple expansion. Zhu designed the FGFA [14] framework, which uses an optical flow model at the feature map level to estimate the motion information between adjacent frames. By combining motion information and adjacent frame features to improve the feature response of the current frame to obtain higher quality detection results, the FGFA framework effectively improves the detection effect of video frames affected by motion blur. However, the content of only part of the key frames in the video shows great changes, and the other adjacent frames have a high degree of correlation. It is not necessary to perform feature fusion on each frame of image. Thus, Zhu [15] only uses the deep convolutional network to extract image features in key frames and combines the optical flow network to fuse the motion information between key frames, while the features of non-key frames are obtained by updating the features of the motion part based on the previous key frame features according to the optical flow network. At the same time, the selection of key frames is adaptively decided based on the quality of the feature map, which steadily improves the

detection performance and runs efficiently. Liu [16] combined the ConvLSTM [17] module at the SSD detection feature map level to process spatial and temporal information at the same time and obtained features with higher timing consistency and quality, allowing for improved detection performance. Xiao designed the spatial temporal memory nodule [18] to achieve video object detection. STMM and ConvLSTM are similar and use a two-way recurrent network to process the information of the preceding and following frames at the same time.

2.3. Self-Attention

Self-attention is a mechanism first introduced in [19] for machine translation. Jaderberg proposed spatial transformer networks [20], which implement global scaling, rotation, and other transformations on the feature map, enabling the network to have invariance of scaling, rotation, and other transformations. Hu proposed a squeeze-and-excitation network [21]. By modeling the correlation between the channels in the convolution feature map, each channel was assigned a different importance weight, thereby recalibrating different channel features. Through this channel domain attention mechanism, the network can combine global features to learn to select and improve features that have a more important impact on the current target and suppress less important features, thereby improving the efficiency and performance of the network. He Kaiming proposed non-local neural networks [22], drawing on the method of non-local means filtering in image processing, using the weighted sum of all location features to represent the feature response of the location so as to model long-distance feature dependence.

3. Kinematic Control of CDHRM

3.1. Overview

In order to make more effective use of the temporal information in the video, the video object detection framework based on local attention is as shown in Figure 1. Given a video, each frame is first processed by a CNN like DarkNet53 [23] to extract features. This is followed by the YOLO detector to predict multi-scale object categories and bounding boxes. Aiming at reducing the imbalance of positive and negative samples, we propose the spatial attention module to classify foreground objects on multi-scale feature layers. The local attention sequence model is used after the spatial attention module to obtain the distribution of spatial attention with temporal consistency, thereby improving the performance of video target detection. We also modify ConvGRU to effectively establish the temporal information across frames, providing higher quality features for the followed detector. In the following sections, we describe in detail the proposed spatial attention, local attention, and modified ConvGRU.

3.2. Spatial Attention

The spatial attention module obtains the distribution of spatial attention by modeling the correlation of features at various positions in space and instructs the model to pay more attention to the areas of the feature that can perform subsequent tasks more effectively. As shown in Figure 2, in order to obtain the spatial attention distribution, firstly, the maximum feature response and the average feature response in each grid of the feature map are obtained through max pooling and average pooling operations. Then, these two features are stitched together, and a small convolutional network is used to model the feature correlation in its local area. Following this, the attention distribution of each area is obtained on this basis. Finally, the attention distribution and the original feature map are multiplied in the spatial dimension to strengthen the more noteworthy regional features, and the remaining unimportant regional features are suppressed.

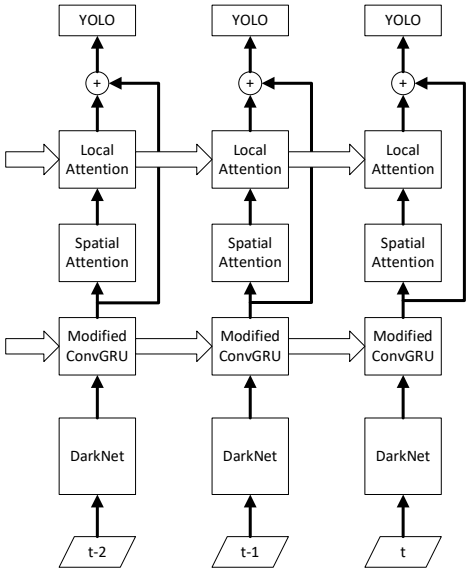

Figure 1. Pipeline of the proposed video object detection framework. Only three frames are shown for simplicity.

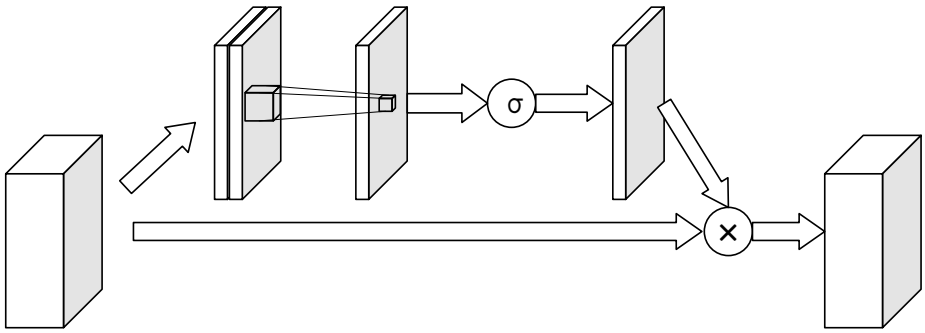

Figure 2. Spatial attention module.

Formally, let $\mathbf{F}$ and $\mathbf{F}'$ be the original feature map and spatial attention feature map, respectively. After obtaining the maximum and average feature response, we use a small convolutional network to compute the distribution of spatial attention as

$$\begin{aligned} \mathbf{F}_{avg} &= \mathrm{AvgPool}(\mathbf{F}) \\ \mathbf{F}_{\max} &= \mathrm{MaxPool}(\mathbf{F}) \\ \mathbf{F}' &= \mathbf{F} \otimes \sigma\left(\mathbf{W}_2 \mathrm{g}\left(\mathbf{W}_1\left[\mathbf{F}_{avg}; \mathbf{F}_{\max}\right]\right)\right) \end{aligned} \tag{1}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are the parameters of the two-layer convolutional network with 3×3 kernel size. Function g represents the ReLU (rectified linear unit) activation function, and σ represents the sigmoid activation function.

3.3. Local Attention Sequence Model

The distribution of spatial attention changes with time and has a certain continuity, and optical flow networks are usually used to establish motion information. However, this requires the introduction of additional deep convolutional networks such as FlowNet, leading to excessive model parameters and calculations, which is not conducive to model deployment. In addition, the optical flow network establishes local pixels corresponding to motion information, and it is difficult to model the continuity between high-level semantic features. Since the motion between adjacent moments occurs more in the local domain, we design the local attention sequence model, focusing on the small-range motion

information in the local domain, so as to establish the temporal consistency of the spatial attention module.

Specifically, the local attention sequence model can be formulated as follows: The first step is to achieve the aligned distribution of spatial attention by aggregating the corresponding feature cells with correspondence weights. Given two adjacent frames $\mathbf{F}_t$ and $\mathbf{F}_{t-1}$, we first compute the affinity between two feature cells at various positions. Then, we compute the normalized correspondence weights in the local area. Finally, we compute the weighted sum of the corresponding feature cells in the local area as the aligned distribution as

$$\begin{aligned} C_{x,y}(i,j) &= \mathbf{F}_t(x,y) \cdot \mathbf{F}_{t-1}(x+i, y+j) \\ \mathbf{T}_{x,y}(i,j) &= \frac{\exp\left(C_{x,y}(i,j)\right)}{\sum\limits_{a,b\in\{-k,\ldots,k\}} \exp\left(C_{x,y}(a,b)\right)} \\ \hat{\mathbf{A}}_t(x,y) &= \sum_{i,j\in\{-k,\ldots,k\}} \mathbf{T}_{x,y}(i,j) \cdot \mathbf{A}_{t-1}(x+i, y+j) \end{aligned} \tag{2}$$

where $C_{x,y}$ represents the affinity matrix. $\mathbf{T}_{x,y}$ represents the correspondence weight matrix, and it is restricted in the sub-region with stride k. $\hat{\mathbf{A}}_t$ represents the aligned distribution of spatial attention.

After achieving the aligned distribution of spatial attention, a small neural network, named the update network, is devised to fuse two distributions adaptively, with the goal of incorporating the temporal context of videos. As shown in Figure 3, the update network takes the concatenation of two distributions to obtain an adaptive weight through a single convolutional network. Then, we compute the weighted sum of the two distributions as the final distribution.

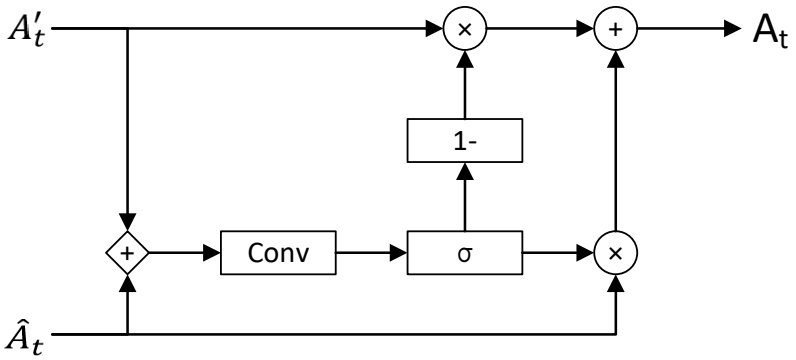

Figure 3. Update network.

3.4. Modified ConvGRU

Video object detection has a high demand for real-time performance. We introduce ConvGRU to establish video temporal information. However, ConvGRU has a large amount of parameters and calculations, which seriously affects the efficiency of the model. Therefore, we modify the traditional ConvGRU and optimize its parameters and calculation so as to improve the video object detection performance without excessively increasing the running time.

The network details of the modified ConvGRU are shown in Figure 4. First, a concatenate layer is used to connect the input state X_t and the hidden state H_{t-1}, which can make full use of the spatial and temporal information. However, it also causes the feature dimension to be expanded by 2 times. Therefore, 1×1 convolution is used to compress the feature dimensions in order to reduce the calculation amount of subsequent modules, and, then, the design of grouped convolution [24] is adopted to optimize the parameters and calculation.

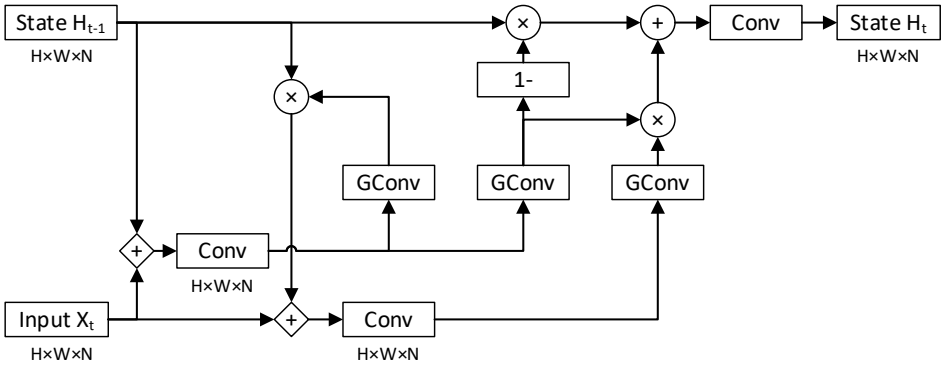

Figure 4. Modified ConvGRU.

Formally, let X and H be the current feature map and the hidden feature map, respectively. Here we use ReLU6 as the activation function to make the gating unit activation sparse so that more unimportant historical information will be forgotten. We establish the temporal information as

$$\begin{aligned} \hat{\mathbf{X}}_t &= \mathbf{W}_1 * [\mathbf{X}_t, \mathbf{H}_{t-1}] \\ \mathbf{Z}_t &= \tfrac{1}{6}\mathrm{ReLU6}\big(\mathbf{W}_z * \hat{\mathbf{X}}_t\big) \\ \mathbf{R}_t &= \tfrac{1}{6}\mathrm{ReLU6}\big(\mathbf{W}_r * \hat{\mathbf{X}}_t\big) \\ \hat{\mathbf{M}}_t &= \mathbf{W}_2 * [\mathbf{X}_t, \mathbf{R}_t \odot \mathbf{H}_{t-1}] \\ \hat{\mathbf{H}}_t &= \mathrm{ReLU}\big(\mathbf{W}_h * \hat{\mathbf{M}}_t\big) \\ \mathbf{H}_t &= \mathbf{W}_3 * \big((1 - \mathbf{Z}_t) \odot \mathbf{H}_{t-1} + \mathbf{Z}_t \odot \hat{\mathbf{H}}_t\big) \end{aligned} \tag{3}$$

where **Z** and **R** represent the update gate and reset gate, respectively; [x,y] is the concatenate operation; $*$ is the convolution calculation; and $\odot$ is the matrix element multiplication. $W_{1\sim3}$ represents 1×1 convolution parameters, and $\mathbf{W}_{\{\mathbf{z},\mathbf{r},\mathbf{h}\}}$ represents grouped convolution parameters. The gating unit activation function uses the ReLU6 function, which makes the gating unit activation sparse, allowing more unimportant historical information to be forgotten.

4. Experiments

4.1. Dataset and Setup

We evaluated our framework on the ImageNet VID [25] dataset, which contains objects of 30 classes with fully annotated bounding boxes. Experiments were based on the PyTorch framework to implement the video object detection model on the basis of local attention. The hardware environment of the training server is Intel Xeon (Skylake) Platinum 8163 2.5 GHz CPU, 32 GB DDR4 memory, NVIDIA V100 16 GB GPU.

The video object detection model training consisted of two stages: first, images were used to train the object detection network that did not contain the sequence model, and, then, the sequence model was introduced to train using video sequences. We used data enhancement with random scaling and cropping, and it was controlled within 1/5 of the image size. The exposure and saturation of the image were randomly adjusted, controlled within 1.5 times on the HSV color space, and finally horizontally flipped randomly with a probability of 50%. During training, we used an SGD optimizer, the momentum coefficient was 0.9, the batch size was 64, the initial learning rate was 0.001, and the weight decay rate was 0.0005. Furthermore, using the warm-up strategy, the learning rate linearly increased from 0.0001 to the initial learning rate in the first 2000 iterations, and then, the learning rate decreased by 10 times at the 40,000th iteration, giving a total of 60,000 iterations.

4.2. Results

We compared our methods with the original YOLO method for video object detection. The results are shown in Table 1, where mAP is the mean average precision metric, P is the precision score, and R is the recall score. F1 = 2*P*R/(P + R).

Table 1. Performance of our method on ImageNet VID.

		mAp	**P**	**R**	**F1**	**Airplane**	**Antelope**	**Bear**	**Bike**	**Bird**	**Bsus**
YOLO		33.77	48.03	28.55	35.81	36.62	11.46	22.62	36.93	31.97	56.55
LA		35.57	55.23	29.43	38.4	70.28	25.34	15.22	42.06	41.35	47.95
car	cattle	dog	cat	elephant	fox	panda	hamster	horse	lion	lizard	monkey
20.34	46.66	8.26	48.47	29.98	47.34	44.85	79.73	55.04	3.99	5.19	0.65
21.92	40.02	3.68	35.99	26.53	50.33	50.01	83.36	44.45	1.95	14.93	0.78
moto	rabbit	redpanda	sheep	snake	squirrel	tiger	train	turtle	boat	whale	zebra
55.48	10.57	10.47	70.11	20.14	0.3	15.77	93.81	34.74	63.91	13.42	37.82
49.16	13.03	18.74	71.27	25.31	0.54	17.56	89.73	38.01	72.37	21.15	34.11

Through the comparison of experimental results, the following conclusions can be drawn: (1) The video target detection based on the local attention sequence model proposed in this section can effectively improve the performance of video target detection. (2) Compared with single-image target detection, mAP is increased by 1.8 points after the introduction of the sequence model, which is a relative increase of 5.3%.

In order to further understand the improvement of the reliability of video object detection by introducing the local attention sequence model, we randomly selected some video object detection results, as shown in Figure 5. It can be seen from the comparison of the results that the video object detection model after the introduction of the local attention sequence model can better solve the difficult detection problems caused by the occlusion of the object movement process in the video, posture transformation, and the blurring caused by camera movement.

4.3. Ablation Study

We conducted an ablation study on ImageNet VID to validate the effectiveness of the proposed modules. The results are shown in Table 2, where mAP is the mean average precision score, time represents the inference time, FLOPs is the calculation of the module, and Parameters represents the number of module parameters.

By comparing the results of ablation experiments, the following conclusions can be drawn: (1) Modified ConvGRU and local attention sequence models can improve the performance of video target detection. (2) Compared with the traditional ConvGRU, the improved ConvGRU increases the amount of parameters and calculations to a lesser extent, with relative increases of 17.3% and 9.5%, respectively, which are equivalent to 24.8% and 30.3%, respectively, of the traditional ConvGRU, and the model accuracy is basically the same. (3) By contrast, the improvement of the performance of ConvGRU is more obvious. mAP increased by 1.09, a relative increase of 3.2%, and the local attention sequence model increased mAP by 0.69, a relative increase of 2.1%. (4) However, the numbers of parameters and calculations added to the local attention sequence model are relatively small, increasing by 1.4% and 3.1%, respectively. (5) Using the modified ConvGRU and local attention sequence model at the same time can improve the performance more obviously, as mAP increased by 1.8 with this method. We randomly select some illustration results of the local attention as shown in Figure 6.

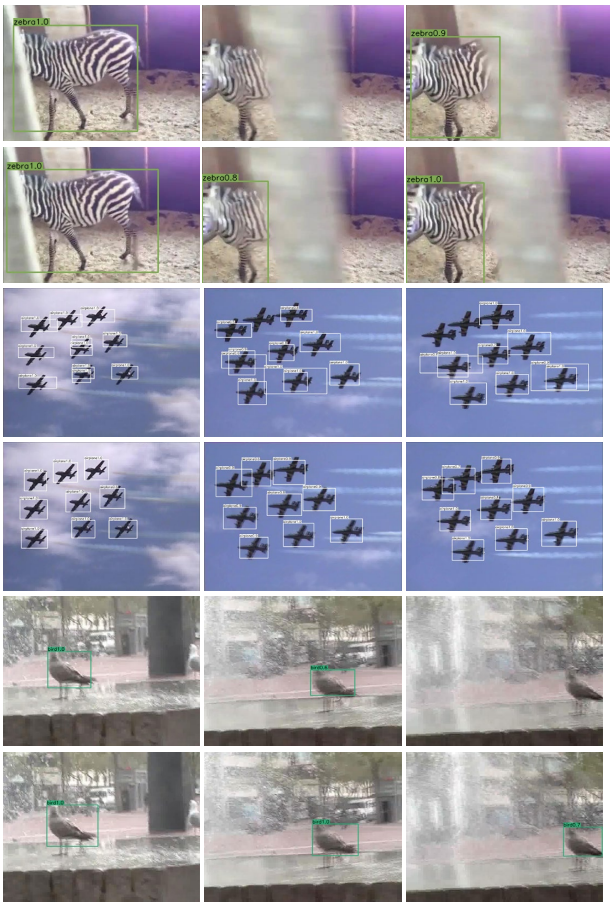

Figure 5. Detection samples.

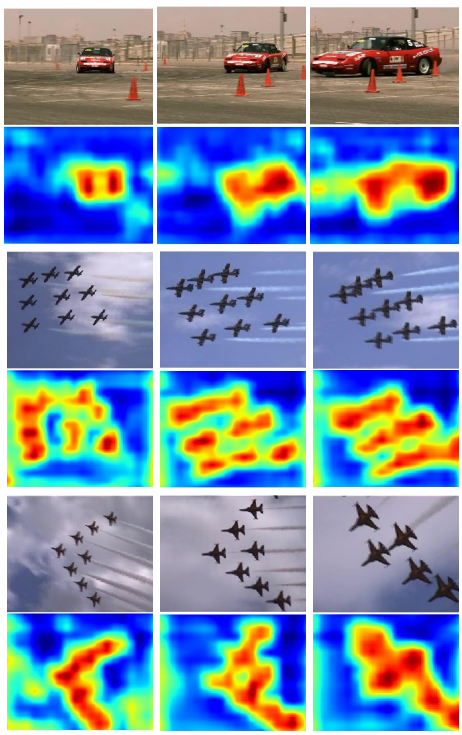

Figure 6. Illustration of local attention.

Table 2. Ablation results.

	YOLO				
ConvGRU		✓			
Modified ConvGRU			✓		✓
Spatial & Local Attention				✓	✓
mAP	33.773	35.026	34.867	34.467	35.571
time/ms	4.731	6.701	5.636	5.019	5.882
FLOPs/G	0.39	0.512	0.427	0.402	0.439
Parameters/M	1.013	1.718	1.188	1.027	1.202

5. Conclusions

In this paper, we examined the video object detection technology that integrates local attention, mainly focusing on three aspects: (1) we designed a spatial attention module to improve the efficiency and accuracy of object detection; (2) we designed a local attention sequence model to process video context and temporal information more efficiently and to solve the problem of low abnormality detection performance in videos; (3) we modified the ConvGRU to more effectively establish temporal information, thereby improving the quality of the video features. We conducted ablation studies on ImageNet VID to examine the effectiveness of our framework in video object detection. The proposed framework achieved 35.57% mAP on ImageNet VID. However, there remains a lack of studies focusing on this topic, and there are similar areas worthy of further research, such as combining video key frames for calculation optimization and introducing additional supervision signals to improve the accuracy of attention distribution.

Author Contributions: Conceptualization, X.Z. and Z.L.; methodology, X.Z.; software, X.Z.; validation, Z.L., H.W. and Y.N.; formal analysis, H.W.; investigation, X.Z.; resources, Y.N.; data curation, Y.N. and J.T.; writing—original draft preparation, X.Z.; writing—review and editing, X.Z.; visualization, Z.L. and H.W.; supervision, Y.N. and Z.L.; project administration, H.W., J.T. and Y.N.; funding acquisition, Z.L., Y.N. and J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the Open Research Project of Technology and Equipment of Rail Transit Operation and Maintenance Key Laboratory of Sichuan Province (No. 2020yw001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
2. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *arXiv* **2019**, arXiv:1905.05055.
3. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, 1137–1149. [CrossRef] [PubMed]
5. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
6. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Häusser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; Brox, T. Flownet: Learning Optical Flow with Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
7. Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; Brox, T. Flownet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2462–2470.

8. Huang, B.; Huang, H.; Lu, H. Convolutional Gated Recurrent Units Fusion for Video Action Recognition. In *International Conference on Neural Information Processing*; Springer International Publishing: Cham, Switzerland, 2017; pp. 114–123.
9. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
10. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.
11. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
12. Kang, K.; Li, H.; Yan, J.; Zeng, X.; Yang, B.; Xiao, T.; Zhang, C.; Wang, Z.; Wang, R.; Wang, X.; et al. T-CNN: Tubelets with Convolutional Neural Networks for Object Detection From Videos. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *28*, 2896–2907. [CrossRef]
13. Han, W.; Khorrami, P.; Paine, T.L.; Ramachandran, P.; Babaeizadeh, M.; Shi, H.; Li, J.; Yan, S.; Huang, T.S. Seq-NMS for Video Object Detection. *arXiv* **2016**, arXiv:1602.08465.
14. Zhu, X.; Wang, Y.; Dai, J.; Yuan, L.; Wei, Y. Flow-Guided Feature Aggregation for Video Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 408–417.
15. Zhu, X.; Dai, J.; Yuan, L.; Wei, Y. Towards High Performance Video Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7210–7218.
16. Liu, M.; Zhu, M. Mobile Video Object Detection with Temporally-Aware Feature Maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5686–5695.
17. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *2015*, 802–810.
18. Xiao, F.; Lee, Y.J. Video Object Detection with an Aligned Spatial-Temporal Memory. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 485–501.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In *Advances in Neural Information Processing Systems*; The MIT Press: Red Hook, NY, USA, 2017; pp. 6000–6010.
20. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. *arXiv* **2015**, arXiv:1506.02025.
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
22. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
23. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
24. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
25. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]