information

Article

Improved Feature Extraction and Similarity Algorithm for Video Object Detection

Haotian You, Yufang Lu * and Haihua Tang

School of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China
* Correspondence: luyufang@glut.edu.cn

Abstract: Video object detection is an important research direction of computer vision. The task of video object detection is to detect and classify moving objects in a sequence of images. Based on the static image object detector, most of the existing video object detection methods use the unique temporal correlation of video to solve the problem of missed detection and false detection caused by moving object occlusion and blur. Another video object detection model guided by an optical flow network is widely used. Feature aggregation of adjacent frames is performed by estimating the optical flow field. However, there are many redundant computations for feature aggregation of adjacent frames. To begin with, this paper improved Faster RCNN by Feature Pyramid and Dynamic Region Aware Convolution. Then the S-SELSA module is proposed from the perspective of semantic and feature similarity. Feature similarity is obtained by a modified SSIM algorithm. The module can aggregate the features of frames globally to avoid redundancy. Finally, the experimental results on the ImageNet VID and DET datasets show that the mAP of the method proposed in this paper is 83.55%, which is higher than the existing methods.

Keywords: video object detection; faster RCNN; feature pyramid; similarity algorithms; dynamic region aware convolution

Citation: You, H.; Lu, Y.; Tang, H. Improved Feature Extraction and Similarity Algorithm for Video Object Detection. *Information* **2023**, *14*, 115. https://doi.org/10.3390/info14020115

Academic Editor: Alessandra Lumini

Received: 22 December 2022
Revised: 2 February 2023
Accepted: 6 February 2023
Published: 12 February 2023

1. Introduction

Recently, many scholars began to pay attention to video object detection. Video object detection has many applications in real scenarios, such as security monitoring, unmanned driving, the internet of things, and intelligent navigation [1,2]. Different from the good progress of image object detection and a lot of applications [3,4], video object detection remains to be studied. Traditional detection methods are mainly divided into Region proposals extraction, feature extraction, classification and other steps. Firstly, the traditional methods select the proposal regions by sliding windows and other methods. Then, feature extraction and classification of region proposals are carried out. Common features include Histogram of Oriented Gradient, Local Binary Pattern, etc. Common classifiers include Support Vector Machine, Naive Bayesian Classifier. Deep learning has shown a strong ability to represent image illumination and other features, leading the research in the field of computer vision. At present, there are two object detection methods, which are two-stage and one-stage models. The R-CNN series is a typical representative of two-stage detection models with high accuracy. In the first stage of the detection process, the region proposal is extracted and the object and background are initially divided. In the second stage, the features of the corresponding region proposals are extracted, and the object location is corrected and the category is predicted. Region Proposal Network was proposed by Faster RCNN to integrate the network structure, which further improved the accuracy and efficiency. Compared with the two-stage model, the one-stage detection is completed in one stage, which has the characteristics of simple structure and high detection efficiency. Compared to images, videos have a high degree of redundancy. There are many problems when the static image detection model is directly applied to video object

detection. Because the objects in the video are constantly changing, and these changes have an impact on the performance of the detection. It is the key point to solving the problem of video object detection. At present, slow-moving objects are relatively easy to detect, but fast-moving objects are difficult to detect accurately. Moving objects are fuzzy and anamorphic. Therefore, it is necessary to aggregate the features of multiple frames. There are many methods for video object detection, mainly including algorithms based on motion information and algorithms based on detection and tracking [5–9]. Deep feature flow for video recognition (DEF) is the first paper to use the concept of key frame in the field of video object detection [10]. It is considered that adjacent frames have similar features, which leads to a large number of features being calculated repeatedly. Kang proposed a tubelet proposal network (TPN), which uses static image object detection combined with a long short-term memory network (LSTM) for video object detection [11]. Zhao trained the model with an SSD object detection frame combined with adjacent frames [12]. Deep learning has made great progress in the application of video object detection, including renewed detection paradigms, datasets, and so on [13–17]. Wu proposed the SELSA module, which aggregates features based on semantics [18]. Only using semantic similarity to cluster images is not comprehensive, so this paper uses the modified SSIM algorithm and feature maps to improve the similarity algorithm and proposes an improved S-SELSA module. The proposed method can compare the similarity more comprehensively and reduce the risk of clustering error without increasing too many redundant calculations. This paper also uses feature pyramid and Dynamic Region Aware Convolution (DRConv) to enhance the feature extraction ability of Faster RCNN. Finally, ImageNet VID and DET datasets were used in the experiment. The method proposed in this paper achieves an mAP of 83.55. Experimental results show that the proposed method has better performance.

2. Faster RCNN

Faster RCNN consists of four main modules. The first module is conv layers, which can output a feature map. Faster RCNN object detection model proposes an RPN network model that is different from RCNN, SPPNet, and Fast RCNN [19]. The RPN is the second module, which can share convolutional layers with the detection models and generate proposals. The third module is ROI Pooling. It collects the input feature maps and proposals from RPN to extract the proposal feature maps, which are sent to the subsequent fully connected layer to classify. It implements end-to-end detection and improves the accuracy of the model. The structure of Faster RCNN is shown in Figure 1. The last module is used for classification and regression. The input is the proposal feature map obtained from the previous layer. The output is the class of the object and the exact location in the image. First, it proposes RPN, which achieves object detection performance with high accuracy. In addition, compared with other one-stage networks, two-stage networks are more accurate. Especially for high-precision, multi-scale and small object problems, the advantages of a two-stage network are more obvious. Faster RCNN works well on multiple datasets and object tasks and often achieves better results after fine-tuning. Finally, there are many points that can be optimized in the whole algorithm framework of Faster RCNN, which provides a broad space for algorithm optimization.

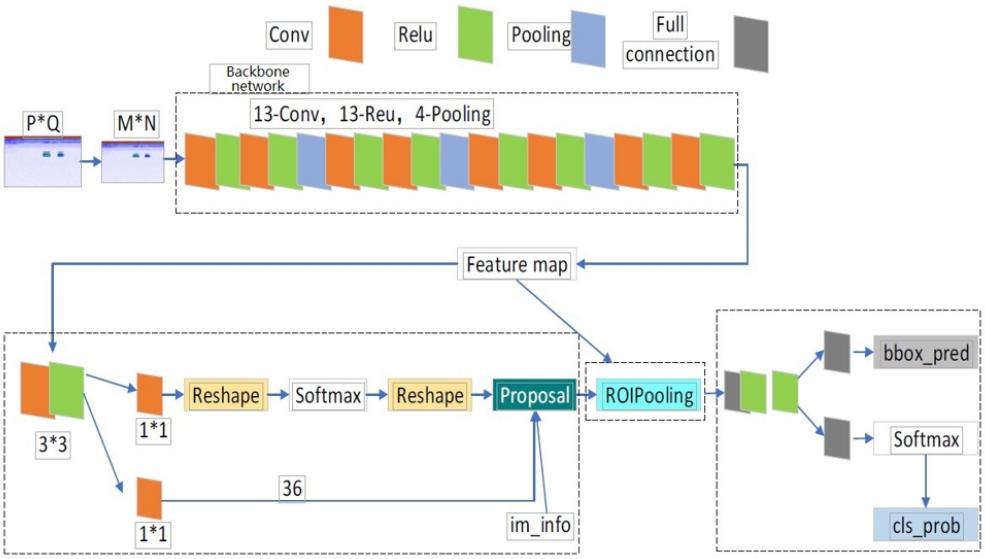

Figure 1. The structure of Faster RCNN.

3. Improved Faster RCNN

Faster RCNN has a large number of parameters. Furthermore, it is prone to overfitting. In the convolution process, small objects are easy to be lost and the recognition effect is bad. Thus, this paper fuses Feature Pyramid (FPN) and ResNet101 as the backbone network and replaces all 3×3 standard convolutions in the last three stages of ResNet101 with DRConv [20–22]. The improved ResNet101 is shown in Figure 2.

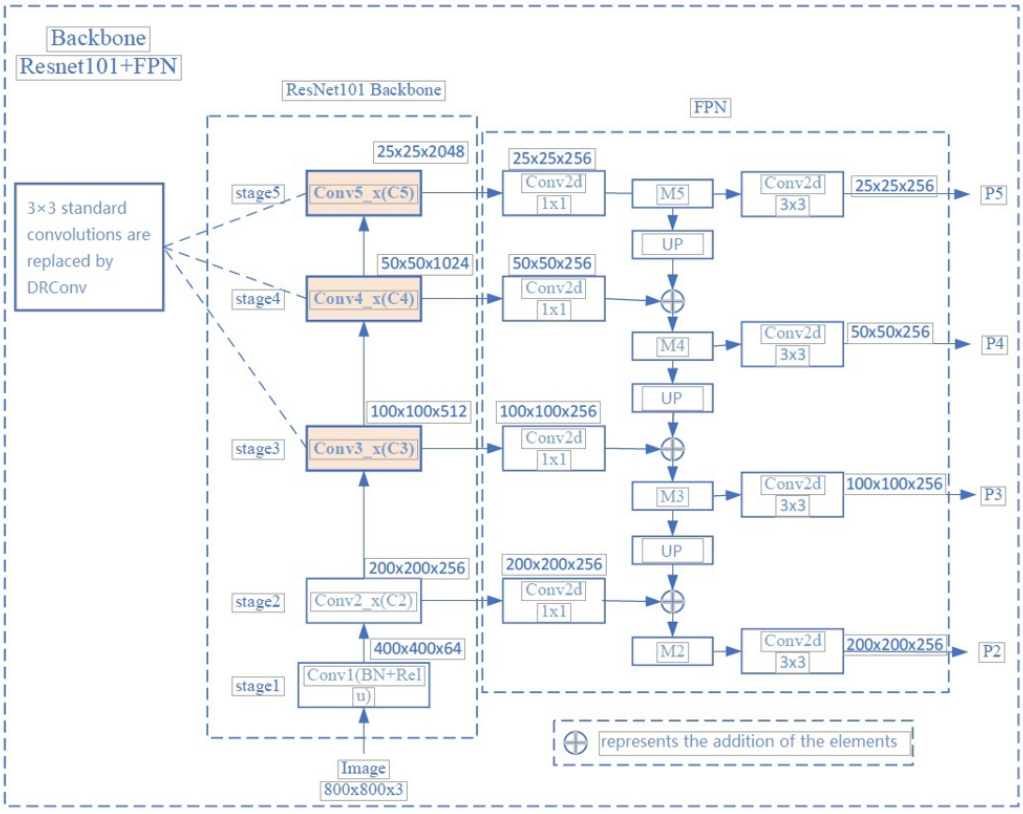

Figure 2. Improved ResNet101.

The current mainstream convolution operations are performed across spatial domains in a filter-shared manner, so more effective information can only be captured when these convolution operations are repeatedly applied. DRConv uses learnable guided masks to transfer the increased filters to the spatial dimension, which not only improves the expressiveness of convolutions but also maintains the computational cost and translation invariance of standard convolutions. $M = \{S_0, \ldots S_{m-1}\}$ is a guided mask that expresses the spatial regions. $W = [W_0, \ldots, W_{m-1}]$ are the filters of regions. The filter $W_t \in \mathbb{R}^C$ is corresponding to the region S_t. The o-th channel of the out feature map is defined as

$$Y_{u,v,g} = \sum_{c=1}^{C} X_{u,v,c} * W_{t,c}^{(o)} \quad (u,v) \in S_t, \tag{1}$$

where $X \in \mathbb{R}^{U \times V \times C}$ is the input of standard convolution and $Y \in \mathbb{R}^{U \times V \times C}$ is the output. $W_{t,c}^{(o)}$ is the c-th channel of $W_t^{(o)}$. The distribution is decided by a learnable guided mask which is a significant module. $M_{u,v}$ can be calculated by

$$M_{u,v} = argmax\left(\hat{F}_{u,v}^{0}, \ldots \hat{F}_{u,v}^{m-1}\right), \tag{2}$$

where $F_{u,v}$ represents the guided feature of each position (u,v). $argmax(.)$ can calculate the maximum value's subscript.

4. Motivation and Method

4.1. Motivation

Fast-moving objects are difficult to be accurately detected due to defocus and occlusion. Furthermore, the adjacent frames in a short period of time are redundant, the calculation is large and the effect of the feature aggregation is not ideal. While aggregating features from multiple frames is an effective approach, aggregating features from just adjacent frames is redundant. From the perspective of similarity algorithms, it is more effective to aggregate features globally. In this paper, the similarity algorithm is considered from two aspects: semantic similarity and SSIM. The semantic similarity and SSIM of each frame in the video are compared, and several frames with the highest similarity are selected to aggregate features instead of temporally adjacent frames.

4.2. Improved Similarity Algorithm

RPN can produce the proposals $X^f = \left\{X_1^f, X_2^f, \ldots\right\}$ of each frame f. From a semantic point of view, the similarity between two proposals $\left(X_i^k, X_l^j\right)$ can be calculated by

$$\omega_{ij}^{kl} = \phi\left(X_i^k\right)^T \psi\left(X_j^l\right). \tag{3}$$

$\phi(.)$ and $\psi(.)$ are two transformation functions. After the ROIPooling of Faster-RCNN, the similarity between two proposal feature maps can be calculated by modified SSIM which is expressed as

$$S_{i,j}^{k,l}(SSIM) = SSIM\left(F_i^k, F_j^l\right) = \frac{\left(2\mu_i\mu_j + k_1^2D^2\right)\left(2\sigma_{ij} + k_2^2D^2\right)}{\left(\mu_i^2 + \mu_j^2 + k_1^2D^2\right)\left(\sigma_i^2 + \sigma_j^2 + k_2^2D^2\right)}. \tag{4}$$

The original SSIM mainly compares pixels of the original image, while the modified SSIM compares pixels of the feature maps. μ_i and μ_j are the averages of all pixes in F_i^k and F_j^l. $F^k = \left\{F_1^k, F_2^k, \ldots\right\}$ are the proposal feature maps of the frame k. σ_i^2 and σ_j^2 are the variances of all pixels. σ_{ij} is the covariance. k_1 and k_2 are often set to 0.01 and 0.03

according to [23]. The value of D can be obtained by subtracting the smallest pixels from the largest pixels [24]. The final similarity can be expressed as

$$\omega_{ij}^{kl} = \phi\left(X_i^k\right)^T \psi\left(X_j^l\right) \frac{\left(2\mu_i\mu_j + k_1^2D^2\right)\left(2\sigma_{ij} + k_2^2D^2\right)}{\left(\mu_i^2 + \mu_j^2 + k_1^2D^2\right)\left(\sigma_i^2 + \sigma_j^2 + k_2^2D^2\right)}. \tag{5}$$

The similarity algorithm can guide the feature aggregation across different proposals. The aggregated feature includes more information. The softmax function is used to normalize the similarity across different proposals. The new feature is expressed as

$$\bar{X}_i^k = \sum_{l\in\Omega}\sum_{j=1}^{N} \omega_{ij}^{kl} X_j^l. \tag{6}$$

where Ω includes all frames used to aggregate features. Figure 3 reveals the architecture with S-SELSA.

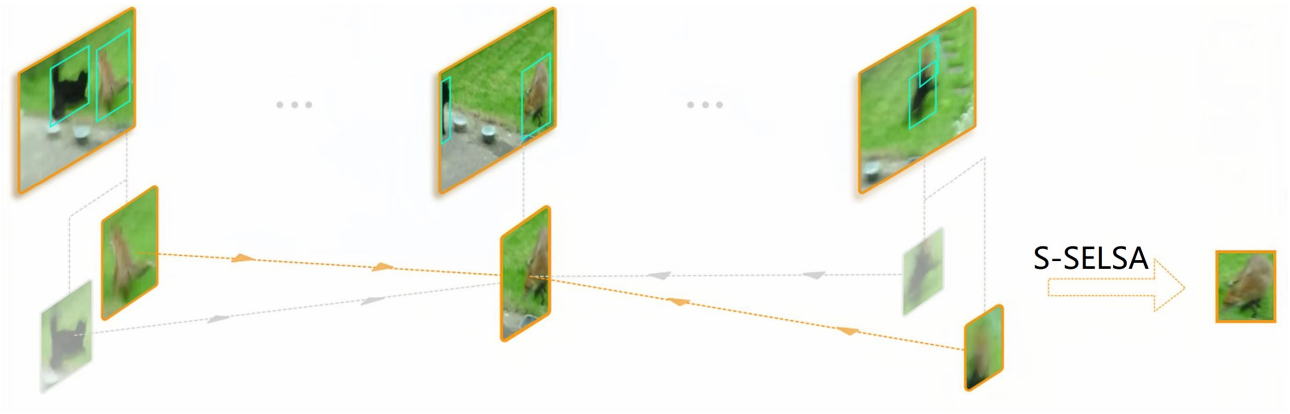

Figure 3. The architecture with S-SELSA. Firstly, some approximate frames are selected from the global perspective through the similarity algorithm proposed in this paper. The module then aggregates the features of these frames.

4.3. A Spectral Clustering Viewpoint

The work process of S-SELSA is closely related to the spectral clustering algorithm. $G = (X, W)$ is a similarity graph on the proposals, where X are nodes and W are edges. Normalizing each row in W to sum 1 can generate the stochastic matrix T. It controls the random walk on G. The transition probability from proposal i and to j is expressed by T_{ij}. Proposals of the same class form a subgraph. The probability of false feature aggregation should be as minimal as possible. The transition probability $P_{\bar{A}A}$ from subgraph $\bar{A} = X - A$ to subgraph A is expressed as

$$P_{\bar{A}A} = \frac{\sum_{i\in\bar{A}, j\in A} \pi_i T_{ij}}{\sum_{i\in\bar{A}} \pi_i}, \tag{7}$$

where π_i denotes the degree of correlation proposals. According to [25], the transition probability is equivalent to the normalized minimum cut,

$$NCut\left(A, \bar{A}\right) = P_{A\bar{A}} + P_{\bar{A}A}. \tag{8}$$

If the optimal partition A is found, T is minimized. The optimization of T is further propagated to the proposal features.

5. Experiments

All experiments used the same environment with 32 GB RAM, GPU NVIDIA GeForce 2080, and a 2 TB hard drive. The experiments use ImageNet VID and DET datasets to train and test different models or methods.

5.1. Datasets

In this paper, the training sets of ImageNet VID and ImageNet DET datasets are used to jointly train the model. ImageNet VID is a video object detection dataset. The training set includes 3862 video clips and the validation set has 555 video clips. The frame rate of each video clip is either 25 or 30 frames per second. Each image frame in the video is annotated and the whole dataset is annotated with 30 object categories. The ImageNet DET dataset is an image object detection dataset whose training set contains 456,567 images and 200 categories. The categories in the ImageNet VID dataset are subsets of the ImageNet DET dataset. Therefore, the images in the ImageNet DET dataset corresponding to the categories of the ImageNetVID dataset were used for training.

5.2. Index of Evaluation

The experiment uses mean average precision(mAP) as an evaluation index to analyze different models or methods. Average precision (AP) is based on the recall ratio and precision ratio. When the degree of confidence is greater than the threshold, the example is positive. Otherwise, it is a negative example. Recall ratio r is the proportion of the number of positive examples correctly detected to the actual total number of positive examples and is expressed as

$$r = \frac{TP}{TP + FN} \tag{9}$$

where TP represents the number of positive examples correctly detected. FN denotes the number of examples that are actually positive but detected as negative. The sum of TP and FN represents the total number of actual positive examples. Precision ratio p denotes the proportion of positive examples to total positive examples in test results, it is can be calculated by

$$p = \frac{TP}{TP + FP} \tag{10}$$

where FP denotes the number of actually negative examples but detected as positive examples. The total number of examples is n. The accuracy ratio is defined as

$$AP = \sum_{k=1}^{n} p_k(r_{k+1} - r_k). \tag{11}$$

The mAP is defined as

$$mAP = \frac{1}{Q} \sum_{q=1}^{Q} AP_q, \tag{12}$$

where Q is the total Categories.

5.3. Implementation Details and Sampling Strategies for Feature Aggregation

The backbone network is improved ResNet101 proposed in this paper. RPN is utilized on the output of conv4. Anchors of 3 scales and 3 aspect ratios are used. Then the input of Fast RCNN is the output of conv5. Two S-SELSA modules are inserted after each fully connected layer. Then there are the classification and boundary box regression modules. $\phi(.)$ and $\psi(.)$ are instantiated as one fully-connected layer. ROIPooling is used to generate proposal feature maps. SSIM algorithm can calculate the degree of similarity by features. Finally, the degree of similarity can be measured by semantics and features. The batch size is 4. The strategy of learning rate decay is adopted. SGD training has 220 k iterations. The learning rate is set to 2.5×10^{-4}. The learning rate is divided by 10 at 110 k and 165 k.

Two random frames in the video are sampled with the corresponding training frame. The images are all adjusted to the shorter side of 600 pixels. It is important for video detection to sample the frames. Better results can be produced by feature aggregation with more frames. Furthermore, the improvements in testing performance need an even stride to sample frames. 21 frames are the number of aggregated frames. The sampling stride is 10. Other methods which use an optical flow or RNN do not work well when the stride is large.

5.4. Ablation Study

As shown in Table 1, the difference of overlap threshold has an impact on the mAP of the models. a 50% overlap threshold provides good performance of different models. In order to better analyze the results, all objects are divided into slow, medium, and fast objects. The moving speed of the object is divided according to the average intersection-over-union(IoU) of the current object and its corresponding object in the neighbouring frames. The smaller the average value of IoU is, the faster the object moves. The average IoU is greater than 0.9 for slow objects and less than 0.7 for fast objects. As shown in Table 2, the experiments test the performance of different models including the model proposed in this paper. The model improved by FPN and DRConv achieves an mAP of 74.21. Compared with the original ResNet101, the performance of the model in this paper has a bit improvements in all experiments. The improvements are mainly due to the better feature extraction capability. The last three experiments show that the S-SELSA module leads to a large 9.93 mAP improvement compared with the baseline. It should be noticed that the mAP (fast) receives the biggest improvements over the baseline. Therefore, the model proposed in this paper has better detection capability for fast-moving objects and has a little boost of 2.33 mAP compared to the model with the SELSA module.

Table 1. Precision of each model with different thresholds.

Positive Overlap Threshold (%)	50	60	70	80
ResNet101	73.62	71.58	68.64	65.69
ResNet101 + FPN	73.88	71.52	67.59	64.33
ResNet101 + DRConv	73.71	70.11	67.58	63.49
ResNet101 + DRConv + FPN	74.21	71.31	69.39	64.98
ResNet101 + SELSA	80.25	77.89	72.54	70.12
ResNet101 + DRConv + FPN + SELSA	81.22	76.99	72.55	70.11
ResNet101 + DRConv + FPN + S-SELSA	83.55	80.08	75.37	71.04

Table 2. The experiment results on the Image VID dataset.

Models	mAP (%)	mAP (%) (Slow)	mAP (%) (Medium)	mAP (%) (Fast)
ResNet101	73.62	82.12	70.96	51.53
ResNet101 + FPN	73.88	82.52	71.41	52.07
ResNet101 + DRConv	73.71	82.58	71.63	52.15
ResNet101 + DRConv + FPN	74.21	82.94	72.05	52.68
ResNet101 + SELSA	80.25	86.91	78.94	61.38
ResNet101 + DRConv + FPN + SELSA	81.22	87.78	79.76	62.15
ResNet101 + DRConv + FPN + S-SELSA	83.55	90.17	82.39	64.78

The experiments use semantic similarity and feature similarity instead of temporal neighbours. Therefore, the experiments sample evenly from a complete video sequence. The method proposed in this paper is useful because it does not rely on any time information (such as optical flow) and does not perform cross-frame feature alignment operations. The method proposed in this paper does not need inaccurate time information to predict and can aggregate features from the entire video sequence.

5.5. Comparison with other Popular Methods

In this paper, other methods have their experiments. Batch gradient descent with added momentum term is used to train the models. The momentum coefficient is set to 0.9. In the experiment, the thresholds of the distance between the nearest frame and the current frame are set as T1 = 3 and T2 = 10. Complete training of the model with all the data of the training set is called epoch training. In the first stage of model training, the training sets of ImageNetDET and ImageNet VID are used for training. The experiments set the data size of each batch as 2 images, and the initial learning rate was 5.0×10^{-4}. The model consists of four iterations. After the first 2 iterations, the learning rate is reduced to 5.0×10^{-5}, and the model parameters are saved. Then stage 2 model training is performed. For the training of FlowNet-SD, the amount of data in each batch is set to 2 images, and the initial learning rate is 4.0×10^{-5}. After 1.333 iterations of the model, the learning rate was reduced to 4.0×10^{-6}. There are two iterations, and then the model parameters are saved. For the training of FlowNetS, the amount of data in each batch is set to 2 images, and the initial learning rate is 2.0×10^{-5}. After 1.333 iterations of the model, the learning rate was reduced to 2.0×10^{-6}. There are two iterations, and then the model parameters are saved. The experiment results are shown in Table 3.

The proposed model is compared with TCN, TPN + LSTM, D(&T loss), and FGFA on the ImageNet VID validation set [26,27]. Among them, the results of TCN, TPN + LSTM, and D((&T loss) are directly provided by the original authors, and the results of FGFA are obtained by rerunning the code provided by the authors under the same training and validation set Settings. As shown in Table 3, the average precision (AP) and mAP of each model on each class of objects include 30 categories such as airplanes, antelope, and bear. The mAP of the proposed model reaches 83.55%, which is 36.05%, 15.15%, 7.35%, and 7.75% higher than that of the TCN, TPN + LSTM, D(&T loss), and FGFA, respectively. Among them, TCN and TPN + LSTM belong to post-processing methods. Compared with these methods, the proposed method clusters frames based on semantic and feature similarity, and each frame is selected from a global perspective. There is a lot of redundancy in frames that are in short succession. The proposed method can fully aggregate the features of each similar frame, and then improve the detection accuracy. The runtime of the method in this paper is 25.6 fps, which is faster than FGFA. The runtime of FGFA is 1.08 fps. Table 4 shows that the mAP of the proposed method on slow, medium, and fast objects is increased by 4.86%, 6.53%, and 9.06%, respectively, compared with the FGFA model. Therefore, the performance of the proposed method is stronger, especially for fast object detection.

Figure 4 shows one result of the object detection in the video by the method in this paper. The results show the blur and occlusion when the object is moving, and the method in this paper can aggregate the features of multiple frames, and finally accurately detect the object.

Figure 4. The detection result of the proposed method.

Table 3. Comparison of AP and mAP of each method on the ImageNet VID dataset.

Object	AP of TCN (%)	AP of TCN + LSTM (%)	AP of FGFA (%)	AP of D (&T Loss) (%)	AP of Our Method
airplane	72.7	84.6	88.1	89.4	90.3
antelope	75.5	78.1	85.0	80.4	88.4
bear	42.2	72.0	82.5	83.8	89.8
bicycle	39.5	67.2	68.1	70.0	76.6
bird	25.0	68.0	72.8	71.8	73.8
bus	64.1	80.1	82.3	82.6	84.2
car	36.3	54.7	58.6	56.8	75.8
cattle	51.1	61.2	71.7	71.0	84.1
dog	24.4	61.6	73.3	71.8	81.2
domestic cat	48.6	78.9	81.5	76.6	83.5
elephant	65.6	71.6	78.0	79.3	82.1
fox	73.9	83.2	90.6	89.9	92.1
giant panda	61.7	78.1	82.3	83.3	91.2
hamster	82.4	91.5	92.4	91.9	93.8
horse	30.8	66.8	70.3	76.8	85.2
lion	34.4	21.6	66.9	57.3	79.2
lizard	54.2	74.4	79.3	79.0	82.3
monkey	1.6	36.6	53.9	54.1	69.6
motorbike	61.0	76.3	84.3	80.3	85.1
rabbit	36.6	51.4	66.7	65.3	78.1
red panda	19.7	70.6	82.2	85.3	87.3
sheep	55.0	64.2	57.2	56.9	74.8
snake	38.9	61.2	74.7	74.1	82.1
squirrel	2.6	42.3	56.5	59.9	76.5
tiger	42.8	84.8	91.0	91.3	92.3
train	54.6	78.1	82.4	84.9	85.7
turtle	66.1	77.2	80.2	81.9	83.2
watercraft	69.2	61.5	65.7	68.3	82.3
whale	26.5	66.9	75.6	68.9	82.2
zebra	68.6	88.5	91.3	90.9	93.7
mAP	47.5	68.4	76.3	75.8	83.55

Table 4. Comparison between FGFA and the proposed method.

Method	mAP(%)(Slow)	mAP(%)(Medium)	mAP(%)(Fast)
FGFA	85.31%	75.86%	55.72%
Our method	90.17%	82.39%	64.78%

6. Conclusions

This paper improves the method of feature extraction model and similarity algorithm to detect objects in videos. The improved method involves replacing all the 3×3 traditional convolutions in the last three stages of ResNet101 with DRConv and incorporating feature pyramids. In terms of similarity algorithm, frames are clustered by considering both semantic and feature similarity. Compared with the method of adjacent frames such as optical flow, this paper uses the modified SSIM algorithm to calculate the feature similarity through the proposal feature maps. Then S-SELSA was proposed from a global perspective to extract the features of frames across time and space by combining semantics and feature similarity to reduce redundancy. Compared with the SELSA module which only clusters from the semantic perspective, the clustering error rate is less and the detection accuracy is higher. Finally, the experimental results on the ImageNet VID dataset show that the mAP of the method proposed in this paper is 83.55%. The mAP of the proposed model reaches 83.55%, which is 36.05%, 15.15%, 7.35%, and 7.75% higher than that of TCN, TPN + LSTM, D(&T loss), and FGFA, respectively. The following research will focus on the improvement of the clustering algorithm to improve the accuracy and detection speed.

Author Contributions: Conceptualization, H.Y., Y.L. and H.T.; methodology, H.Y., Y.L. and H.T.; formal analysis, H.Y. and Y.L.; investigation, H.Y. and Y.L.; data curation, H.Y. and H.T.; writing—original draft preparation, H.Y. and H.T.; writing—review and editing, H.Y. and H.T.; supervision H.Y., Y.L. and H.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Guangxi Key Laboratory of Embedded Technology and Intelligent System,Guilin University of Technology(CN) grant number (2020-2-7).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

S-SELSA	Structural Similarity SELSA
DEF	Deep Feature Flow for Video Recognition
TCN	Temporal Convolutional Network
TPN	Tubelet Proposal Network
LSTM	Long Short Term Memory Network
FGFA	Flow-Guided Feature Aggregation for Video Object Detection
D&T	Detect to Track and Track to Detect
SSD	Single Shot MultiBox Detector
DRConv	Dynamic Region Aware Convolution
FPN	Feature Pyramid Network
RPN	RegionProposal Network
SGD	Stochastic Gradient Descent

References

1. Yan, H.; Huang, J.; Li, R.; Wang, X.; Zhang, J.; Zhu, D. Research on video SAR moving target detection algorithm based on improved faster region-based CNN. *J. Electron. Inf. Technol.* **2021**, *43*, 615–622.
2. Du, L.; Wei, D.; Li, L.; Guo, Y. SAR target detection network via semi-supervised learning. *J. Electron. Inf. Technol.* **2020**, *42*, 154–163.
3. Zhang, Y.; Cai, W.; Fan, S.; Song, R.; Jin, J. Object Detection Based on YOLOv5 and GhostNet for Orchard Pests. *Information* **2022**, *13*, 548. [CrossRef]
4. Wang, J.; Yu, L.; Yang, J.; Dong, H. DBA SSD: A novel end-to-end object detection algorithm applied to plant disease detection. *Information* **2021**, *12*, 474. [CrossRef]
5. 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.
6. Zhu, H.; Wei, H.; Li, B.; Yuan, X.; Kehtarnavaz, N. A review of video object detection: Datasets, metrics and methods. *Appl. Sci.* **2020**, *10*, 7834. [CrossRef]
7. 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.
8. Han, W.; Khorrami, P.; Paine, T.L.; Ramachandran, P.; Babaeizadeh, M.; Shi, H.; Huang, T.S. Seq-nms for video object detection. *arXiv* **2016**, arXiv:1602.08465.
9. Wang, S.; Zhou, Y.; Yan, J.; Deng, Z. Fully motion-aware network for video object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
10. Zhu, X.; Xiong, Y.; Dai, J.; Yuan, L.; Wei, Y. Deep feature flow for video recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
11. Kang, K.; Li, H.; Xiao, T.; Ouyang, W.; Yan, J.; Liu, X.; Wang, X. Object detection in videos with tubelet proposal networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
12. Zhao, B.; Zhao, B.; Tang, L.; Han, Y.; Wang, W. Deep spatial-temporal joint feature representation for video object detection. *Sensors* **2018**, *18*, 774. [CrossRef] [PubMed]
13. Girshick, R. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 7–13 December 2015.

14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
16. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Ferrari, V. The open images dataset v4. *Int. J. Comput. Vis.* **2020**, *128*, 1956–1981. [CrossRef]
17. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014.
18. Wu, H.; Chen, Y.; Wang, N.; Zhang, Z. Sequence level semantics aggregation for video object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
21. Lin, T.Y.; 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, Nashville, TN, USA, 20–25 June 2021.
22. Chen, J.; Wang, X.; Guo, Z.; Zhang, X.; Sun, J. Dynamic region-aware convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
23. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
24. Wang, Z.; Liu, X.; Huang, L.; Chen, Y.; Zhang, Y.; Lin, Z.; Wang, R. Model pruning based on quantified similarity of feature maps. *arXiv* **2021**, arXiv:2105.06052.
25. Meilă, M.; Shi, J. A random walks view of spectral segmentation. *Int. Workshop Artif. Intell. Stat.* **2001**, *2001*, 203–208.
26. Kang, K.; Ouyang, W.; Li, H.; Wang, X. Object detection from video tubelets with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
27. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Detect to track and track to detect. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.