applied sciences

MDPI

Article

Temporal-Guided Label Assignment for Video Object Detection

Shu Tian *,†, Meng Xia † and Chun Yang

School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

* Correspondence: shutian@ustb.edu.cn

† These authors contributed equally to this work.

Abstract: In video object detection, the deterioration of an object's appearance in a single frame brings challenges for recognition; therefore, it is natural to exploit temporal information to boost the robustness of video object detection. Existing methods usually utilize temporal information to enhance features, often ignoring the information in label assignments. Label assignment, which assigns labels to anchors for training, is an essential part of object detection. It is also challenged in video object detection and can be improved by temporal information. In this work, a temporal-guided label assignment framework is proposed for the learning task of a region proposal network (RPN). Specifically, we propose a feature instructing module (FIM) to establish the relation model among labels through feature similarity in the temporal dimension. The proposed video object detection framework was evaluated on the ImageNet VID benchmark. Without any additional inference cost, our work obtained a 0.8 mean average precision (mAP(%)) improvement over the baseline and achieved a mAP(%) of 82.0. The result was on par with the state-of-the-art accuracy without using any post-processing methods.

Keywords: video object detection; temporal-guided label assignment; feature instructing module

Citation: Tian, S.; Xia, M.; Yang, C. Temporal-Guided Label Assignment for Video Object Detection. *Appl. Sci.* **2022**, *12*, 12314. https://doi.org/10.3390/app122312314

Academic Editor: Andrea Prati

Received: 21 October 2022
Accepted: 28 November 2022
Published: 1 December 2022

1. Introduction

Recently, object detection has made significant progress by relying on the fast development of deep convolutional neural networks. However, object detection may fail due to the deteriorated appearance of an object caused by issues such as motion blur, video defocus, part occlusion, and rare poses in videos. Video object detection has emerged as a new area in computer vision to address these problems.

Video contains inherent temporal information, which can boost the performance of video object detection. Previous efforts in leveraging temporal information focused on features. Some works exploit image-level feature information from other frames to improve detection. Deep feature flow (DFF) [1], flow-guided feature aggregation (FGFA) [2], and towards high performance (THP) [3] rely on external optical flow [4] guidance to compute motion among frames on feature maps to boost performance or improve speed. Others [5–7] are end-to-end without any guidance networks and may be more computationally efficient or faster. Some other works have been conducted on the instance-level features after the region proposal network. Sequence level semantics aggregation (SELSA) [8] selects proposal features globally. Relation distillation networks (RDN) [9] distillates instance-level features gradually using multiple stages. Hierarchical video relation networks [10] seek richer semantic features, which are not limited to a certain video or local area. Temporal ROI align (TROI) [11] proposes an alternative operator with temporal information for the traditional region of interest (ROI) component; however, these methods only operate on features, often ignoring the information in label assignments.

Label assignment is an essential part of object detection in training. Traditional label assignment methods resort to manually predefined rules to sample the positive and

negative locations. Some methods, such as RetinaNet [12] and the fully convolutional one-stage object detector (FCOS) [13], rely on prior knowledge such as intersection over union (IOU) and ground truth (gt) box locations for sampling. These methods rely on prior human knowledge and cannot adapt to a variety of objects; therefore, dynamic label assignment strategies have been proposed. GuidedAnchoring [14] and MetaAnchor [15] dynamically change the anchor set during training. Feature selective anchor-free (FSAF) [16] method presents an anchor-free branch for guiding label assignment; however, these methods still somehow depend on prior knowledge. There are other methods that are completely data-driven. Optimal transport assignment (OTA) [17] defines label assignment as an optimal transfer problem (OT) from a global perspective. AutoAssign [18] treats all the locations inside the gt boxes as positives and assigns weights for these locations. In AutoAssign, the weights are the labels. Data-driven methods learn from data to adapt to the variance of objects for label assignment; however, this kind of method is also challenged in video object detection. These labels usually fail or underperform when there is a deterioration in the appearance of objects. As shown in Figure 1, frame v is a blurred image, and the foreground and background are easily confused, so its positive labels are also unfocused. Frame $v^{'}$ is a clear image whose positive labels are focused. The unfocused labels are not conducive to model training (the model has no bias towards positive and negative samples).

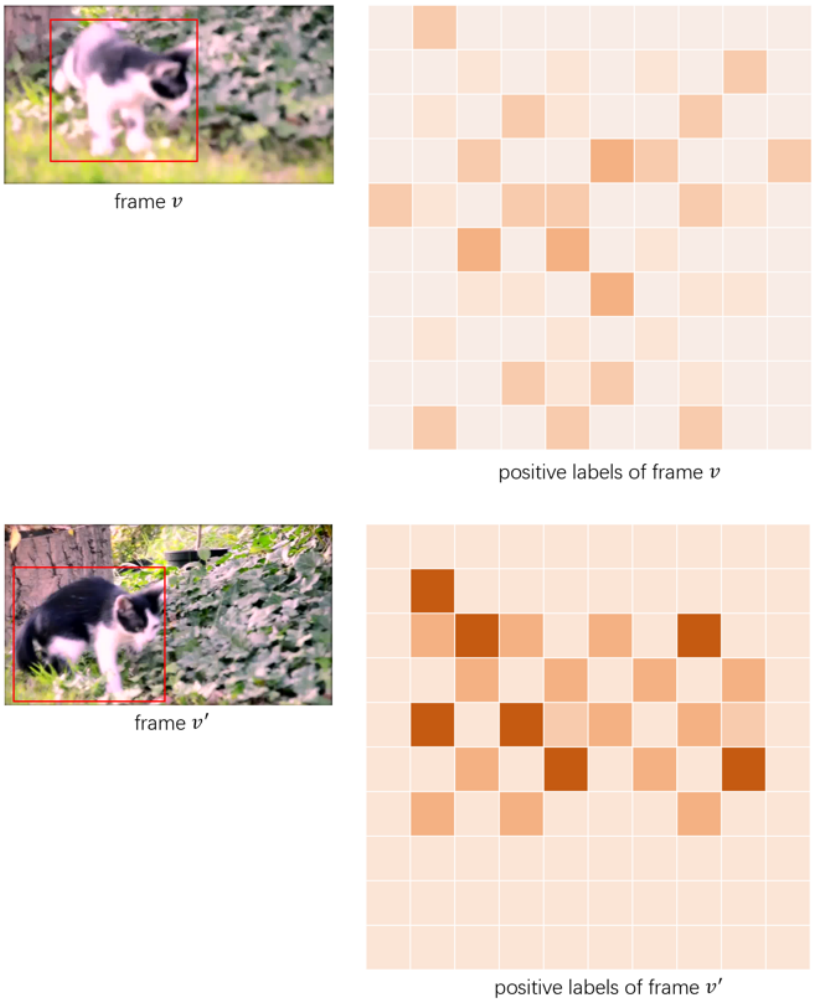

Figure 1. The visualization of the positive labels produced by our baseline (without temporal information, which will be introduced in the experiment) for two frames. Frame v and frame $v^{'}$ belong to the same snippet. We only show the positive labels inside the gt box, and the positive labels outside the gt box are all 0 (i.e., the samples outside the gt box cannot be positive samples).

Specifically, we can resort to temporal information to deal with this issue. We operated on labels generated by label assignment in the same way as features, i.e., we enhanced labels in the temporal dimension. When an object's appearance deteriorates, the labels can be adjusted using the labels from other frames containing normal objects; this can correct erroneous or poorly focused labels. Our work proposes a temporal-guided label assignment framework that stabilizes the learning process of region proposal network (RPN). Specifically, we put forward a feature instructing module (FIM), which utilizes the similarity of features at different locations to establish the relationship of the corresponding labels.

In summary, our contributions are three-fold:

- A temporal-guided label assignment framework is proposed. This framework utilizes temporal information to improve the learning task of an RPN.
- We introduce a feature instructing module (FIM) that additionally uses feature information to establish the temporal relation model of labels and enhances the labels. Features contain richer semantic information and can be used to effectively construct the association of labels.
- Experimental results on the ImageNet VID dataset demonstrate that our proposed framework is on par with state-of-the-art works with no extra inference cost compared with our baseline.

2. Related Work

2.1. Object Detection in Still Images

Object detection in still images is a fundamental task in multimedia and computer vision. The current state-of-the-art object detectors can be divided into two categories: two-stage and one-stage object detectors. The two-stage object detector is a series of methods based on a region proposal (region-based convolutional neural networks (RCNN) [19], fast-RCNN [20], faster-RCNN [21], region-based fully convolutional networks (R-FCN) [22], etc.). The other is the one-stage object detector such as you only look once (YOLO) [23], single shot MultiBox detector (SSD) [24], RetinaNet [12], and FCOS [13], which only uses a convolutional neural network (CNN) to predict the classes and positions of different objects directly.

In the field of video object detection, there are two directions for improving accuracy and speed. Our aim is to improve the accuracy, so we chose the two-stage object detector.

2.2. Video Object Detection

Compared with object detection in still images, methods for video object detection take temporal information into account. To improve performance, feature aggregation of existing state-of-the-art frameworks works on a frame-level or a proposal-level.

Frame-level video object detection (VOD) methods mainly take frames in the same video as context information to enhance the robustness of feature map representation. Early methods often use optical flow as external guidance [1–3]. Instead of relying on optical flow, progressive sparse local attention (PSLA) [5] tackles feature alignment in an end-to-end manner. It establishes the spatial correspondence between the features across frames in the local area through self-attention [25].

On the other hand, proposal-level works operate on boxes. SELSA [8] introduces a semantic aggregation module to fuse the high-level proposal features [26,27]. Gong et al. put forward a Temporal ROI Align operator [11] in place of a traditional ROI align in object detection.

2.3. Label Assignment

Current CNN-based object detectors perform dense detection by classifying anchors and regressing boxes. The process of setting classification and regression targets for anchors is called label assignment.

Traditional label assignment methods usually employ predefined rules to assign ground-truth boxes or backgrounds for anchors. RetinaNet [12] exploits the intersection-

over-union (IOU) ratio to label anchor boxes. FCOS [13] defines the anchor points within a certain center range of the gt as positive samples. Such predefined rules rely on prior human knowledge to allocate samples in space and scale and cannot adapt well to changes in data and the emergence of new instances.

In view of the shortcomings of fixed label assignment strategies, the latest detectors propose adaptive label assignment strategies. GuidedAnchoring [14] dynamically changes the shape of anchor boxes according to the distribution of objects. MetaAnchor [15] randomly samples anchors of any shape during training to cover different types of object boxes. Moreover, in the adaptation of anchor priors, some approaches focus on the sampling method for each object. FSAF [16] embeds an anchor-free branch as a label assignment guide for anchor-based practices.

The label assignment methods mentioned above are still somehow dependent on human knowledge. The latest works bring forward a fully-adaptive strategy for label assignment. AutoAssign [18] takes all points in a gt box as positive samples and then calculates the weights of these sample points. The weights are completely learned from the data, so they are entirely data-driven. OTA [17] believes that AutoAssign still has shortcomings because it only finds the optimal label assignment strategy for a single object and cannot consider context information from a global perspective. Therefore, OTA proposes a global optimal allocation strategy, which defines label assignment as an optimal transfer problem (OT). However, fully data-driven label assignment also cannot adapt well to video object detection when object appearance deterioration occurs in videos. We propose temporal-guided label assignment to enhance the labels. This enables the label assignment to perform well even when the object's appearance deteriorates.

3. Methodology

3.1. Overview

The architecture of our method of temporal guided Label assignment (TGLA) is illustrated in Figure 2. Given a video, while processing a frame it is noted as the target frame I_t. The adjacent frames of the target frame are specified as support frames $\{I_{t'}\}$. First, we proposed a temporal-guided label assignment framework for the learning task of the RPN. Labels produced by label assignment are enhanced. The blue box above is the forward propagation process of the two-stage object detection model. The orange box below is the training process. To establish the labels' relation model, we then introduce a feature instructing module (FIM). All our proposed modules ultimately act on the loss function. Therefore, our method can improve model accuracy without increasing inference costs.

3.2. Temporal-Guided Label Assignment Framework

As shown in Figure 2, the blue box above is the process of forwarding propagation. The image is sent to the backbone to obtain feature maps. Then, the feature maps are sent to the RPN to obtain the classification and coarse regression results. 'Inference' indicates the subsequent forward propagation process, which includes ROI Align and the detection head, to obtain the final detection results. The orange box below is the module for calculating the RPN loss. The label assignment adopts the automatic label assignment strategy in AutoAssign [18]. The traditional form of IOU label assignment is not used, and all locations can be positive and negative samples. We use weight maps to represent the likelihood that a sample is positive or negative.

3.2.1. Weight Maps for Label Assignment

As shown in Figure 2, we follow the initial setting in Faster R-CNN, which means our method is an anchor-based one.

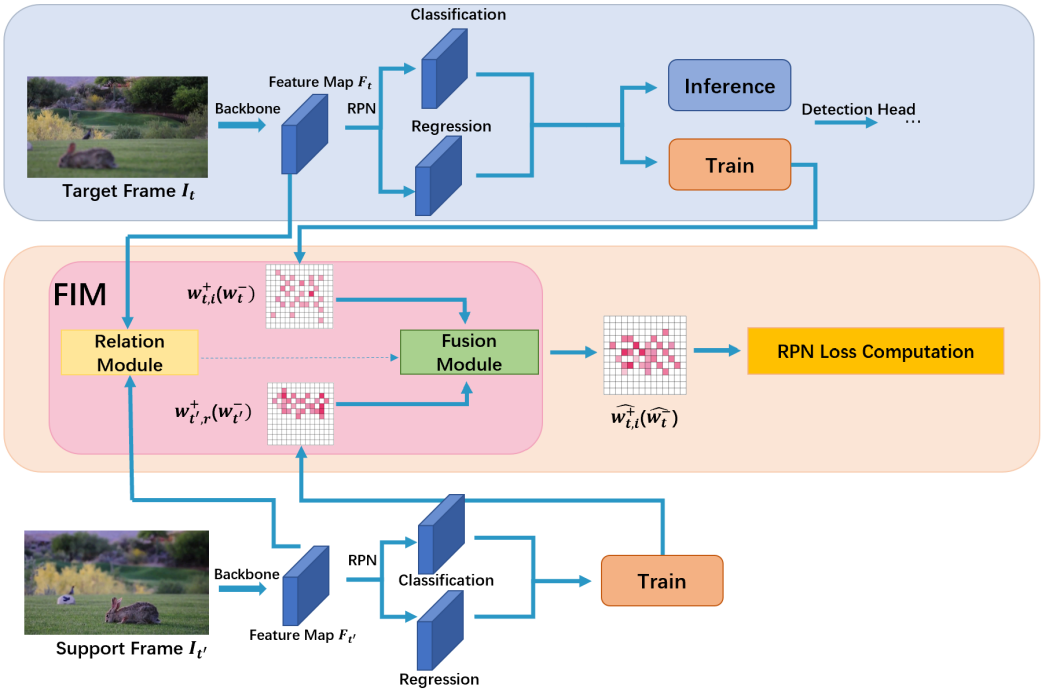

Figure 2. The architecture of our temporal guided label assignment (TGLA). The temporal-guided label assignment framework is presented by the upper blue box and the lower orange box, which denote forward propagation and loss computation separately. Note that the framework we mentioned below does not include feature instructing. Furthermore, we put forward an feature instructing module (FIM), which utilizes feature maps to guide the fusion of weight maps.

w is the weight map of one gt box of one image. Each location's weight represents the likelihood of a positive sample (negative sample). Here, w^+ is the weight map of positives and w^- is the weight map for the negatives; these are generated by center weighting G and confidence weighting C. Confidence weighting is generated by the RPN outputs. We follow the implementation in AutoAssign [18] for our label assignment. For location $q \in S_i$, where S_i denotes all locations inside the bounding box of object i we have the following:

$$w^+(q) = \frac{C(P^+(q))G(\overrightarrow{d(q)})}{\sum_{j \in S_i} C(P^+(j))G(\overrightarrow{d(j)})} \tag{1}$$

where $\overrightarrow{d(q)}$ denotes the offsets of a certain position inside an object to its box center along the x-axis and the y-axis. $P^+(q) = P(cls, q) \cdot P(loc, q)$ is the positive confidence. $P(cls, q)$ and $P(loc, q)$ are the results of classification and coarse regression. Here, $w^+(q)$ is normalized by all candidate locations of one object. $w^-(q)$ is formulated by the IOU of prediction boxes on location q between gt boxes as follows:

$$w^-(q) = 1 - f(iou(q)) \tag{2}$$

where $f(\cdot)$ denotes a function whose role is that the larger the IOU, the less likely the box will be a negative sample, and the smaller it is implies the opposite.

But even this strategy of automatic label assignment encounters challenges in video object detection. When ab object's appearance in the video deteriorates, the label assignment strategy cannot obtain good weight maps because the distinction between the foreground and background is unclear in this condition. For example, the appearance of the rabbit in the target frame in Figure 2 is blurred, and the resulting weight maps are unfocused. Therefore, a framework that utilizes temporal information for the learning task of the RPN is put forward, and we call it temporal-guided label assignment.

3.2.2. Temporal-Guided Label Assignment

First, introducing temporal information to w^+ generates $w^+_{t,i} \in \mathbb{R}^{H\times W}$. $w^+_{t,i}$ is the positive weight map of the i-th gt box of frame t. In the same way, w^-_t is the negative weight map of frame t. H and W are the height and width of $w^+_{t,i}(w^-_t)$ and they are the same as the height and width of the feature maps $F_t \in \mathbb{R}^{H\times W\times C}$ of frame t. H, W, and C denote the height, width, and channel of F_t, respectively. To enhance $w^+_{t,i}(w^-_t)$, we fuse weight maps $\left\{w^+_{t',r}\right\}_{t'\in V, r=g(t,i,t')}(\{w^-_{t'}\}_{t'\in V})$ from other frames into $w^+_{t,i}(w^-_t)$ to form $\widehat{w}^+_{t,i}(\widehat{w}^-_t)$. Our framework is shown in Figure 3:

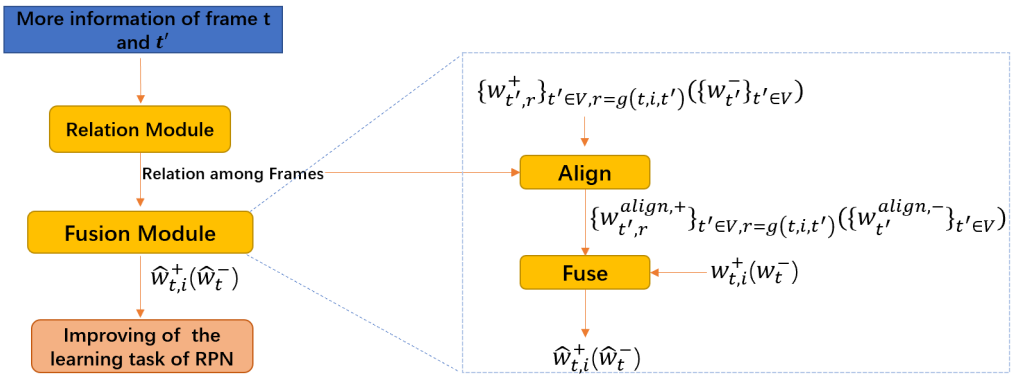

Figure 3. Temporal-guided label assignment framework. 'More information' can be anything that comes from different frames. The fused $\widehat{w}^+_{t,i}(\widehat{w}^-_t)$ act on the learning task of the region proposal network (RPN).

Here, V is the set of the N support frames of frame t, and r is the r-th gt box of frame t' selected by the max IOU with the i-th gt box of frame t. Finally, $\widehat{w}$ plays a significant role in improving the learning task of the RPN. To establish the relation model of the weight maps in a better way, we proposed a feature instructing module.

3.3. Feature Instructing Module

Compared with labels, features contain richer semantic information representing the relationship between different locations. Therefore, we use additional feature information instead of weight maps to model their relationship.

3.3.1. Relation Module

The relation of feature maps F_t and $\{F_{t'}\}_{t'\in V}$ is utilized to guide the fusion of weight maps $w^+_{t,i}(w^-_t)$ and $\left\{w^+_{t',r}\right\}_{t'\in V, r=g(t,i,t')}(\{w^-_{t'}\}_{t'\in V})$. Feature maps of the target frame and support frame set are separately noted as F_t and $\{F_{t'}\}_{t'\in V}$. To calculate the similarity of the two feature maps, we realize the inner product of their corresponding positions by multiplying the two matrices. For each $F_{t'}$ and F_t, the similarity map $Z_{t'}$ can be expressed as follows:

$$Z_{t'} = Z(F^\star_t, F^\star_{t'}) : \mathbb{R}^{(H\times W)\times C} \otimes \left[\mathbb{R}^{(H\times W)\times C}\right]^T \rightarrow \mathbb{R}^{H\times W\times H\times W} \tag{3}$$

where $F^\star$ denotes the feature maps normalized along the channel dimension and $[\cdot]^T$ denotes the transposition of the matrix. $\otimes$ denotes the matrix multiplication. Then, for each location u $\in \{(1,1),(1,2),\cdots,(H,W)\}$ in F_t, we select the most similar K locations in $F_{t'}$ by the similarity map to generate feature point scores $z_{t'}(u,k)$:

$$z_{t'}(u,k) = \frac{Z_{t'}(u,k)}{\sum_{j=1}^{K} Z_{t'}(u,j)} \tag{4}$$

where k denotes one specific location in the top K locations. The feature point scores are calculated by the normalization of the similarity map.

3.3.2. Fusion Module

There is a one-to-one correspondence between the points on the feature map and the points on the weight map. Using the feature point scores $z_{t'}(u,k)$ calculated by Equation (4), for each location u on $w^+_{t,i}(w^-_t)$, each $w^+_{t',r}(w^-_{t'})$ aligned for $w^+_{t,i}(w^-_t)$ can be calculated:

$$w^{align,+}_{t',r}(u) = \sum_{j=1}^{K} z_{t'}(u,j) * w^+_{t',r}(j),\ w^{align,-}_{t'}(u) = \sum_{j=1}^{K} z_{t'}(u,j) * w^-_{t'}(j) \tag{5}$$

where K denotes the top K similar locations to location u in the support frame. $w^{align,+}_{t',r}(w^{align,-}_{t'})$ is extracted by concatenating all $w^{align,+}_{t',r}(u)(w^{align,-}_{t'}(u))$. For N support frames, the N weight maps calculated by Equation (5) are fused with the weight maps of the target frame:

$$\widehat{w}^+_{t,i} = w^+_{t,i} + \sum_{t'\in V, r=g(t,i,t')} \overline{a}^+_{t',r} \cdot w^{align,+}_{t',r},\ \widehat{w}^-_t = w^-_t + \sum_{t'\in V} \overline{a}^-_{t'} \cdot w^{align,-}_{t'} \tag{6}$$

where $\overline{a}^+_{t',r}(\overline{a}^-_{t'})$ denotes the weight score of the fusion, which is calculated by the process in Figure 4 and $\cdot$ denotes element-wise multiplication.

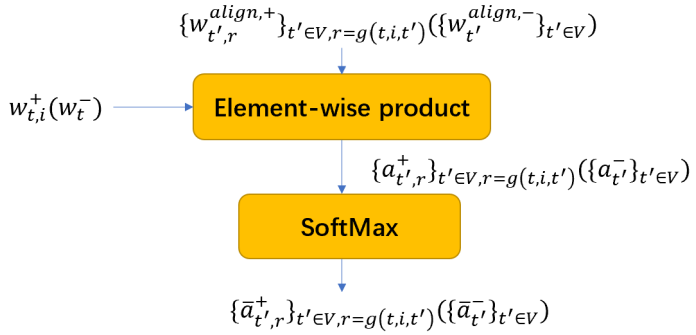

Figure 4. Calculation process of $\left\{\overline{a}^+_{t',r}\right\}_{t'\in V, r=g(t,i,t')}(\{\overline{a}^-_{t'}\}_{t'\in V})$. $a^+_{t',r}(a^-_{t'}) \in \mathbb{R}^{H\times W}$ and $\overline{a}^+_{t',r}(\overline{a}^-_{t'}) \in \mathbb{R}^{H\times W}$.

3.3.3. Improving the Learning Task

The fused weight maps $\widehat{w}$ are generated for the computation of the loss in the RPN to improve the learning task. The loss of positive samples and negative samples are computed separately:

$$L = -\sum_{i=1}^{M} log(\sum_{o\in S_{t,i}} \widehat{w}^+_{t,i}(o) P^+_{t,i}(o)) - \sum_{e\in S} log(1 - \widehat{w}^-_t(e) P^-_t(e)) \tag{7}$$

where $S_{t,i}$ denotes the locations inside the i-th gt box of frame t on the output feature maps and S represents all the locations on the feature maps. M is the gt box number of frame t. P^+ and P^- are the positive and negative confidence, respectively. $P^- = P(cls)$. $\widehat{w}$ is a temporal-guided label generated by our framework, which can perform well in various emergencies such as motion blur, video defocus, part occlusion, and rare poses.

4. Experiment

4.1. Dataset and Evaluation Setup

ImageNet VID [28] is the only mainstream dataset in video object detection, it consists of 3862 video snippets in the training set and 555 snippets in the validation set. There are 30 object categories with fully annotated bounding boxes, which are the subset of the

categories in the ImageNet DET dataset [28]. Following the protocols in [29,30], we trained our model on the 30 overlapped classes of the ImageNet VID and DET sets. We only tuned the parameters and trained the model on the training set. We then reported the mean average precision (mAP), where the IoU threshold for all mAP calculations in our article was 0.5, on the validation of VID.

4.2. Implementation Details

4.2.1. Architecture

We employed a variant of the Temporal ROI Align [11] method called TROI-Var as our baseline. It adopts the label assignment method in [18] for model training. We used ResNet-101 as the backbone for ablation studies and state-of-the-art (SOTA) comparison.

4.2.2. Training Details

We implemented our method mainly on Pytorch [31]. Due to the data richness of the video dataset itself (objects were moving) and the sufficient data volume of the entire dataset, we did not use data augmentation techniques. In both training and inference, the images were resized to the shorter side of 600 pixels. The whole architecture was trained on 4 RTX 2080ti GPUs with SGD. There was one mini-batch in each GPU, and the mini-batch size was 1. What is more, seven epochs of training were performed on our framework, and the initial learning rate was 0.001 and it was divided by 10 at the 5-th and 7-th epoch.

4.3. Ablation Study

4.3.1. Effect of Support Frame Number N

We adopted the uniform sample strategy from the Temporal ROI Align [11] method. That is, the sampling stride is dynamically adjusted according to the length of the video. Table 1 shows the effect of different values of N. When N increased from 2 to 8, the performance increased from 81.0 mAP(%) to 82.0 mAP(%). Compared with the baseline, when N was 2, the accuracy dropped because of the introduction of interference. When N exceeded 8, the accuracy continued to increase within a certain range, but the accuracy no longer increased or decreased after exceeding a certain value. The performance might continue to rise, but we did not continue to verify this due to the limitation of computing resources. The bottleneck was the relation module and fusion module, which were calculated N times in the training phase. Therefore, this did not affect the inference speed. This showed that as the number of support frames increased, the fused weight map was more stable and less prone to large deviations; therefore, N was chosen to be 8 as default.

Table 1. The effect of support frame number N.

#N	2	4	6	8
mAP(%)	81.0	81.4	81.7	82.0

4.3.2. The Effectiveness of Temporal-Guided Label Assignment and the Feature Instructing Module (FIM)

TROI-Var denotes the variant of Temporal ROI Align [11], which introduces the label assignment strategy in AutoAssign [18], and it is the baseline of our model. As shown in Table 2, temporal-guided label assignment only uses the label assignment information to model the relationship between themselves without additional feature information. This improved performance over the baseline, which means the framework is effective. FIM further enhanced the performance of the framework. This demonstrated that the relation model established by features is significant.

Table 2. The effectiveness of the temporal-guided label assignment framework and feature instructing module. Temporal-guided label assignment is the framework without feature instructing, and FIM is the feature instructing module.

Method	Temporal-Guided Label Assignment	FIM	mAP(%)
TROI-Var	-	-	81.2
Our TGLA	✓	-	81.5
Our TGLA	✓	✓	82.0

4.4. Comparison with State-of-the-Art Methods

We proposed a new paradigm for label assignment information aggregation instead of feature aggregation, which is significant and inference cost-free. We compare our work with several recent state-of-the-art approaches on ImageNet VID in Table 3. With bells and whistles, our proposed model can achieve a 82.0 mAP(%) with the ResNet-101 backbone. Our method is on par with most of the SOTA approaches and even outperforms a quantity of them.

Table 3. Performance comparison with other state-of-the-art models on the ImageNet VID validation set. TGLA is our method with the framework and FIM. Here, R101 is ResNet-101 and DCN is a deformable convolutional network.

Method	Backbone	mAP(%)
FGFA [2]	R101	78.4
D&T [32]	R101	80.0
PLSA [5]	R101+DCN	80.0
SELSA [8]	R101	80.25
Leveraging [33]	R101-FPN	81.0
RDN [9]	R101	81.8
TGLA	R101	82.0

4.5. Detection Visualization

Figure 5 showcases one hard example in video object detection. The domestic cat is presented in a blurred appearance during movement. The top two rows show the result of the baseline and our TGLA without FIM. Compared with the result of our TGLA in the third row, they have a lower confidence. This first shows that it is effective to add temporal information to label assignment during training because the result of the second row was better than that of the first row. The result in the third row outperformed that in the second row, which reflects the effectiveness of FIM. This shows that the method of building the relation model of labels also needs to be considered.

Figure 5. Comparison of detection results of different models for hard samples.

5. Conclusions and Outlook

In this work, we proposed a novel framework: temporal-guided label assignment for video object detection. This framework utilizes temporal information to enhance the labels to improve the learning task of the RPN. To better establish the labels' relation model, a novel module, FIM, was proposed in which feature similarity is used to find the association of labels. It is worth mentioning that our method can not only work on two-stage object detectors, but it can also work on single-stage object detectors. We conducted ablation studies on ImageNet VID to prove the effectiveness of our framework on video object detection. Our work obtained a 0.8 mAP(%) improvement over the baseline without any additional inference costs. The proposed framework achieved a 82.0 mAP(%) on ImageNet VID, which is on par with state-of-the-art methods. For future work, it will be interesting to develop techniques for object tracking over frames among multiple objects by incorporating contextual information. On one hand, we have seen that graph matching, with the aid of recent deep matching networks as a backbone, has played a promising role in this direction. On the other hand, it is attractive to adapt temporal prediction models, e.g., the long-term multi-modal tracker, into predictive tracking, especially for the settings whereby the objects' trajectories are relatively well regularized.

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

Funding: This work is supported by the National Key Research and Development Program of China (2020AAA0109701), the National Natural Science Foundation of China (62076024, 62006018), and the Interdisciplinary Research Project for Young Teachers of USTB (Fundamental Research Funds for the Central Universities, FRF-IDRY-21-018).

Data Availability Statement: The data presented in this study are openly available at https://image-net.org/challenges/LSVRC/2017/index.php, accessed on 21 October 2022, reference number [28].

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

References

1. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2349–2358.
2. 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 (ICCV), Venice, Italy, 22–29 October 2017; pp. 408–417.
3. 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 (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7210–7218.
4. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, 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 (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
5. Guo, C.; Fan, B.; Gu, J.; Zhang, Q.; Xiang, S.; Prinet, V.; Pan, C. Progressive sparse local attention for video object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27–28 October 2019; pp. 3909–3918.
6. Jiang, Z.; Liu, Y.; Yang, C.; Liu, J.; Gao, P.; Zhang, Q.; Xiang, S.; Pan, C. Learning where to focus for efficient video object detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 18–34.
7. Bertasius, G.; Torresani, L.; Shi, J. Object detection in video with spatiotemporal sampling networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 331–346.
8. 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 (ICCV), Seoul, Republic of Korea, 27–28 October 2019; pp. 9217–9225.
9. Deng, J.; Pan, Y.; Yao, T.; Zhou, W.; Li, H.; Mei, T. Relation distillation networks for video object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27–28 October 2019; pp. 7023–7032.
10. Han, M.; Wang, Y.; Chang, X.; Qiao, Y. Mining inter-video proposal relations for video object detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 431–446.
11. Gong, T.; Chen, K.; Wang, X.; Chu, Q.; Zhu, F.; Lin, D.; Yu, N.; Feng, H. Temporal ROI align for video object recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Virtual, 2–9 February 2021; Volume 35, pp. 1442–1450.

12. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
13. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27–28 October 2019; pp. 9627–9636.
14. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region proposal by guided anchoring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2965–2974.
15. Yang, T.; Zhang, X.; Li, Z.; Zhang, W.; Sun, J. Metaanchor: Learning to detect objects with customized anchors. In *Advances in Neural Information Processing Systems*; Curran Associates: New York, NY, USA, 2018; Volume 31.
16. Zhu, C.; He, Y.; Savvides, M. Feature selective anchor-free module for single-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 840–849.
17. Ge, Z.; Liu, S.; Li, Z.; Yoshie, O.; Sun, J. Ota: Optimal transport assignment for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 303–312.
18. Zhu, B.; Wang, J.; Jiang, Z.; Zong, F.; Liu, S.; Li, Z.; Sun, J. Autoassign: Differentiable label assignment for dense object detection. *arXiv* **2020**, arXiv:2007.03496.
19. 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 (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
20. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; Curran Associates: New York, NY, USA, 2015; Volume 28.
22. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*; Curran Associates: New York, NY, USA, 2016; Volume 29.
23. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
24. 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.
25. 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*; Curran Associates: New York, NY, USA, 2017; pp. 5998–6008.
26. Zhu, X.; Li, Z.; Lou, J.; Shen, Q. Video super-resolution based on a spatio-temporal matching network. *Pattern Recognit.* **2021**, *110*, 107619. [CrossRef]
27. Zhu, X.; Li, Z.; Li, X.; Li, S.; Dai, F. Attention-aware perceptual enhancement nets for low-resolution image classification. *Inf. Sci.* **2020**, *515*, 233–247. [CrossRef]
28. 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]
29. 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.* **2017**, *28*, 2896–2907. [CrossRef]
30. Lee, B.; Erdenee, E.; Jin, S.; Nam, M.Y.; Jung, Y.G.; Rhee, P.K. Multi-class multi-object tracking using changing point detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 68–83.
31. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*; Curran Associates: New York, NY, USA, 2019; Volume 32.
32. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Detect to track and track to detect. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3038–3046.
33. Shvets, M.; Liu, W.; Berg, A.C. Leveraging Long-Range Temporal Relationships Between Proposals for Video Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27–28 October 2019; pp. 9756–9764.