Spatio-Temporal Learnable Proposals for End-to-End Video Object Detection

Khurram Azeem Hashmi[1,2]
khurram_azeem.hashmi@dfki.de
Didier Stricker[1,2]
didier.stricker@dfki.de
Muhammamd Zeshan Afzal[1,2]
muhammamd_zeshan.afzal@dfki.de

[1] German Research Centre for Artificial Intelligence (DFKI) Kaiserslautern, Germany
[2] Technical University of Kaiserslautern Germany

Abstract

This paper presents the novel idea of generating object proposals by leveraging temporal information for video object detection. The feature aggregation in modern region-based video object detectors heavily relies on learned proposals generated from a single-frame RPN. This imminently introduces additional components like NMS and produces unreliable proposals on low-quality frames. To tackle these restrictions, we present SparseVOD, a novel video object detection pipeline that employs Sparse R-CNN to exploit temporal information. In particular, we introduce two modules in the dynamic head of Sparse R-CNN. First, the Temporal Feature Extraction module based on the Temporal RoI Align operation is added to extract the RoI proposal features. Second, motivated by sequence-level semantic aggregation, we incorporate the attention-guided Semantic Proposal Feature Aggregation module to enhance object feature representation before detection. The proposed SparseVOD effectively alleviates the overhead of complicated post-processing methods and makes the overall pipeline end-to-end trainable. Extensive experiments show that our method significantly improves the single-frame Sparse R-CNN by 8%-9% in mAP. Furthermore, besides achieving state-of-the-art 80.3% mAP on the ImageNet VID dataset with ResNet-50 backbone, our SparseVOD outperforms existing proposal-based methods by a significant margin on increasing IoU thresholds (IoU > 0.5).

1 Introduction

Video Object Detection (VOD) aims to localize and classify objects in a series of subsequent video frames. Recent efforts in video object detection demonstrate that exploiting feature aggregation of temporal information [8, 10, 11, 13, 14, 15, 40, 41, 42, 45, 46] produce superior performance than leveraging temporal information at the post-processing stage [16, 22, 23, 34]. The former approaches mainly enhance the target frame feature representation through aggregating features from neighbouring frames or an entire video clip by designing a specific module, thereby boosting detection results. The majority of these works [15, 21, 35, 40, 42] employ two-stage detectors such as Faster R-CNN [31] or R-FCN [6] to design their VOD pipelines.

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

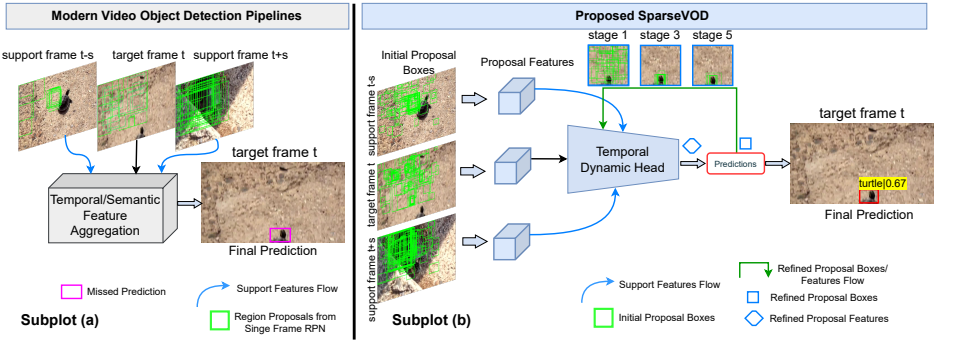

Figure 1: Comparison between previous region-based and the proposed video object detection methods. Subplot (a): Despite the temporal feature aggregation from support frames *t-s* and *t+s* in existing proposal-based approaches, the detector overlooks *Turtle* in pink on a low-quality target frame *t* due to unreliable region proposals. Subplot (b): To tackle such limitations, we propose a novel SparseVOD framework that iteratively refines region proposals through leveraging spatio-temporal feature aggregation prior to final detection. The Temporal Dynamic Head is explained in Fig. 2.

Albeit the enormous success, it is important to highlight that the temporal feature aggregation scheme of all prior region-based VOD methods [6, 14, 15, 21, 40, 42] heavily relies on object proposals from RPN [31] trained without any temporal information. Consequently, these methods suffer from several underlying restrictions. First, they require an additional step of NMS [30] at the beginning to perform dense-to-sparse matching of hand-crafted anchors, making the overall VOD pipeline not end-to-end optimizable. Second, on low-quality frames (appearance deterioration), the generated object proposals are unreliable, leading to ineffective temporal feature aggregation. Fig. 1(a) illustrates that although the detector leverages spatio-temporal information from support frames *t-s* and *t+s*, it fails to detect *Turtle* at the target frame *t*. The main reason for such missed prediction is that generated proposals from single-frame RPN overlooks *Turtle* in the target frame *t*. This corrupts object features during instance-level feature aggregation. The third restriction is that these methods require several support frames to calibrate proposal feature representation for the target frame, decreasing the run time performance. Fourth, since the RPN in these methods is optimized on a single IoU level (generally IoU=0.5), they struggle to provide high-quality detections (IoU > 0.5) despite producing impressive performance on a lower IoU threshold. Moreover, these methods necessitate complex post-processing methods [3, 24] or additional proposal classifier networks [13, 14] to accomplish state-of-the-art performance.

To tackle the aforementioned challenges, we propose SparseVOD, an end-to-end trainable framework that exploits temporal information to learn sparse (merely 100) object proposals for VOD. The SparseVOD employs the recently introduced Sparse R-CNN [37] that has shown impressive performance by eliminating the need for dense priors enumerating over frames and alleviating the interaction between object queries and dense frame features. In particular, motivated from [11, 42], we incorporate a Temporal (Region of Interest) RoI Feature Extraction (TFE) head that replaces a single image RoI extractor in the dynamic instance interactive head in [37]. Furthermore, inspired by [42], we fuse attention guided Semantic Proposal Feature Aggregation (SPFA) module that enhances the feature representation of object proposals in target frames through semantic level sequence aggregation from support frames.

Contrary to most prior VOD works, our SparseVOD operates on a sparse-in sparse-out matching scheme. This not only eliminates components like post-processing and NMS but also enables faster training network convergence without pre-training the detector as done in [19]. Furthermore, thanks to the spatio-temporal learning, iterative refinement of object proposals lead to successful predictions even on a low-quality target frame t as shown in Fig. 1(b). Moreover, our spatio-temporal proposal learning alleviates the need for several support frames in a video and brings significant performance gains on increasing IoU thresholds (see Fig. 4).

Herein, our main contributions are as follows. (1) We propose SparseVOD, a novel end-to-end trainable video object detection method. To our knowledge, *this is the first work that exploits temporal information to learn object proposals for video object detection*. (2) We extend the design of Sparse R-CNN [37] by introducing a Temporal Feature Extraction (TFE) module that leverages temporal information to extract RoI proposal features. Furthermore, we fuse Semantic Proposal Feature Aggregation (SPFA) in [37] to enhance object feature representation before final detection inspired by [11, 42]. (3) By introducing the proposed TFE and SPFA modules in [37], our SparseVOD improves the baseline by far (5-6% mAP). Without bells and whistles, our SparseVOD achieves the *new best mAP of 80.3% on the ImageNet VID benchmark using ResNet-50 as the backbone*. Moreover, it surpasses prior state-of-the-art methods by far in terms of high-quality detections (higher IoU thresholds, see Fig. 4) and achieves optimal speed-accuracy tradeoff (Fig. 5).

2 Related Work

Proposal Learning for Image Object Detection. Ren *et al*. [31] introduce the Region Proposal Network (RPN) in Faster R-CNN to predict object proposals. The RPN consists of a small fully convolutional network [28] that receives an anchor as an input, classifies it as an object or background, and performs box regression. This design is widely incorporated in later two-stage approaches [1, 6, 18, 26]. MetaAnchor [44] proposes to exploit meta-learning to generate anchors dynamically. Cascade RPN [39] improves the object proposal quality of the conventional RPN through multi-stage refinement and adaptive convolution. Recently, Sparse R-CNN [37] introduces a sparse-in sparse-out paradigm that simplifies the sophisticated two-stage object detection pipeline by alleviating complex components such as Non-Maximum Suppression (NMS) [30] and dense priors. Following a similar line of work, this paper proposes spatio-temporal learnable proposals to simplify the video object detection pipeline.

Exploiting Temporal Information in Video Object Detection. Exploiting temporal information from other frames in a video is a natural choice to tackle the challenges of video object detection, and our work derives from the same idea. Existing approaches leveraging temporal information mainly follow one of the two directions. The first line of works [16, 22, 23, 34] mainly employ temporal information to make still-image detection results more coherent and stable. The performance of such VOD methods is sub-optimal because they are not end-to-end trainable and heavily rely on the capabilities of the initial still-image detector. On the contrary, the other direction of methods [8, 10, 11, 13, 14, 15, 20, 40, 41, 42, 45, 46] utilizes temporal information during the course of training. Earlier works in this category adopt FlowNet [9] to propagate warp features across frames [24, 41, 45, 46]. However, temporal exploitation of optical flow-based works is limited to neighbouring frames, yielding inferior performance in occlusions. PSLA [12] proposes to learn the spatial correspondence

between neighbouring frames by employing the progressive sparser stride. All these methods can only capitalize temporal information from a small number of nearby frames to refine the target frame features. Alternatively, global feature aggregation methods [7, 35, 42] have been proposed to utilize long-term semantic information. Recent VOD methods adopt this aggregation scheme and propose blending of temporal features [8], class-aware feature aggregation [13, 14], temporal RoIAlign [11], and temporal meta-adaptor [40] to achieve state-of-the-art results. Although these methods produce superior performance from prior efforts, their feature aggregation rely on object proposals generated without temporal information.

Refining Object Proposals in Video Object Detection. Recent efforts have shown that enhancing object proposal features can alleviate the obstacles of object confusion in videos [15, 21, 35]. Shvets *et al*. [35] refine the proposal for the target frame by learning similarities between proposals from different frames. LSTS [21] models the spatio-temporal correspondence to alleviate misalignment before aggregating features from different frames. Han *et al*. [15] propose integrating inter-video and intra-video proposals to boost target proposal features. *Despite the promising improvements, the effectiveness of all these methods heavily relies on the initial quality of object proposals retrieved from single-frame RPN* [31]. Alternatively, this paper exploits temporal information to generate object proposals for video object detection.

Recently, MAMBA [36] proposes to extract region proposals from the enhanced feature maps through a pixel-level memory bank. TransVOD [19] proposes the transformer-based VOD pipeline by extending Deformable DETR [47] to exploit temporal information in videos. Despite the simple and end-to-end trainable framework, the temporal transformer in TransVOD depends on object queries generated by the spatial transformer optimized without temporal information. Furthermore, owing to the interaction between each object query and the dense features of an entire frame, [19] is not a pure sparse method [37]. As a result of this dense interaction, TransVOD requires pre-training the detector on a similar dataset [25]. On the contrary, our method leverages temporal information to generate object proposals. Moreover, following [37], our proposed SparseVOD operates on a pure sparse paradigm and does not require pre-training the detector.

3 Method

Overview. This section explains our proposed SparseVOD, which consists of two main components: Temporal RoI extraction and Semantic Proposal Feature Aggregation incorporated in the Temporal Dynamic Head. Finally, we discuss network optimization. Note that due to space constraints, the detailed explanation of the still-image object detector Sparse R-CNN [37] is omitted here and can be found in the supplementary material.

3.1 SparseVOD Architecture

The SparseVOD is a simple, end-to-end trainable framework, as shown in Fig. 2. It receives a target frame and multiple support frames from the same video as input and outputs the class and location of objects in the target frame. For each target and support frame, the extracted feature maps from the backbone, proposal boxes and corresponding proposal features are fed into the iterative Temporal Dynamic head consisting of multiple stages. We create Temporal Dynamic Head by incorporating two main components into the dynamic head used in [37] to effectively exploit the temporal information in videos. First, inspired by [11], we leverage support frame RoI features to extract temporal RoI features for the target frame. Then,

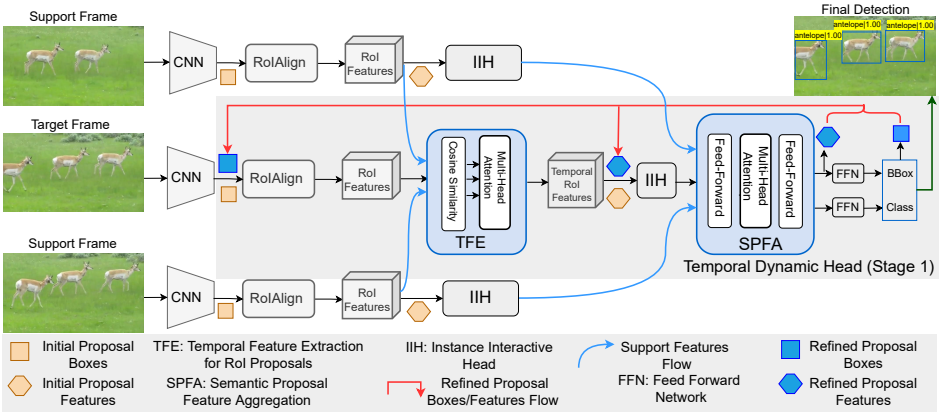

Figure 2: Our SparseVOD Framework. The input for the first stage of the temporal dynamic head consists of initial proposal boxes, proposal features, and spatial features for the target and support frames. The TFE extracts RoI proposal features for the target frame by exploiting RoI features from support frames. Then, SPFA receives object features from each IIH (identical as in [37]) and aggregates object features with guided attention heads. This enhanced object feature representation from SPFA and box predictions after FFN serve as input proposal features and proposal boxes for the next iterative stage.

the corresponding Instance Interactive Head (IIH) produces object feature representation for each video frame. Subsequently, the Semantic Proposal Feature Aggregation module enhances the representation of the target frame by intelligently aggregating object features from support frames. The optimal object features are fed to the corresponding feed-forward network for classification and regression. Similar to [37], we follow the iterative architecture in our SparseVOD. In the next iteration, the newly refined object features and the bounding box predictions serve as the target frame's proposal features and proposal boxes.

Temporal RoI Feature Extraction. The Sparse R-CNN [37] applies the conventional RoIAlign [18] pooling operation to extract proposal features, and it is widely adopted in existing VOD methods [15, 42, 45]. However, the naive RoIAlign operation only restricts the proposal feature extraction to exploit intra-frame features. Therefore, motivated by [11], we incorporate the Temporal Feature Extraction (TFE) module in our Temporal Dynamic Head, as illustrated in Fig. 2. The TFE leverages temporal information for the same object instance across support frames in a video. Since the features of the same object instance have high semantic resemblance across video frames, we calculate cosine similarities between target proposal features and support frame feature maps. Given target proposal features P_t and feature map from support frame F_{t+s}, the cosine similarity $Sim_{t+s}(m)$ is computed as $P_t(m) \otimes F_{t+s}$, where m represents the spatial location of P_t and $\otimes$ denotes dot product. Note that the target frame proposal is mapped on the most similar feature maps from support frames to extract the most similar RoI features. Following the temporal attentional feature aggregation in [11], we adopt the self-attention mechanism [38] to aggregate the RoI features of target and support frames.

Semantic Proposal Feature Aggregation. The Semantic Proposal Feature Aggregation (SPFA) head aims to learn the enhanced feature representation from target RoI features (containing temporal information of the same instance) and support frame RoI features to perform final classification and regression. Since we already have temporal RoI features, we

follow the spirits of [42] and adopt semantic similarity as the metric to aggregate features from support frame proposals. We compute semantic similarity between target and support proposal features in the same way as in TFE. For effective feature aggregation, the SPFA applies multi-head attention [38] on temporal proposal features of a target frame $\overline{P_t}$ and support proposal features P_s as follows:

$$W_t = softmax(\frac{\phi(P_s)\cdot(\theta(\overline{P_t}))^T}{\sqrt{d_{\theta(\overline{P_t})}}})\cdot\sigma(P_s) \tag{1}$$

where $\phi(.)$, $\theta(.)$, and $\sigma(.)$ are some linear transformations. The symbol T represents transposition, d denotes the size of transformed $\overline{P_t}$, and W_t is the enhanced proposal representation of a target frame. This rich instance representation improves robustness against inherent challenges of VOD, such as appearance deterioration.

Loss Function. Since our SparseVOD operates on a one-to-one label matching, our method's loss function and training process are similar to the original Sparse R-CNN. We adopt set predictions loss [2, 19, 47], which aims to optimize the bipartite matching among the ground truth and predictions. Following [2, 19, 37], the cost function is defined as $\mathcal{L} = \lambda_{cls}\cdot\mathcal{L}_{cls} + \lambda_{L1}\cdot\mathcal{L}_{L1} + \lambda_{giou}\cdot\mathcal{L}_{giou}$, where $\mathcal{L}_{cls}$ is the focal loss [27] for classification. $\mathcal{L}_{L1}$ and $\mathcal{L}_{giou}$ are L1 loss and generalized IoU loss [32] for regression, respectively. λ_{cls}, λ_{L1}, and λ_{giou} are coefficients to balance the loss. We employ an identical setting to balance these losses as in [37].

4 Experiments and Results

4.1 Experimental Settings

We perform experiments on the ImageNet VID dataset [33], which comprises 3862 training videos and 555 validation videos. Following prior works [6, 42, 45], we train our model on a combination of ImageNet VID and DET datasets and evaluate the results on the validation set. We adopt ImageNet pre-trained [8] ResNet-50 [17], ResNet-101, and ResNeXt-101 [43] backbones to compare performance with recent state-of-the-art methods. We train our network for 12 epochs with a batch size of 8 on 8 GPUs. Analogous to [37], we use AdamW [29] optimizer with a weight decay of 10^{-4}. Initially, the learning rate is set to 2.5×10^{-5} and divided by 10 at the 8-th and 11-th epochs. Following [2, 37, 47], we set λ_{cls}=2, λ_{L1}=5, and λ_{giou}=2. We follow the basic settings of [37] and set the number of iterative stages, proposal boxes, and the corresponding proposal features to 6, 100, and 100, respectively. We refer readers to supplementary materials for more details.

4.2 Main Results

We compare the performance of the proposed SparseVOD with prior state-of-the-art VOD methods on ImageNet VID dataset in Table 1. Besides the conventional mAPs @IoU=0.5, we compute mAPs @IoU=0.75 and @IOU=0.5:95 as in [25] to analyze the precision of detections. Owing to the unavailability of code at the time of experiments, apart from [19, 36], we reproduce the results of existing methods from the code provided by the original papers for direct comparison. It is important to mention that all the results shown in Table 1 are without any post-processing. By looking at results under the backbone of ResNet-50, it is evident that our SparseVOD outperforms recent methods [4, 11, 42, 45], mainly relying on feature aggregation of region proposals. Furthermore, it surpasses the previous best score of 79.9% by [19] and achieves a new best score of 80.3% on ResNet-50. Note that alongside the

Methods	Venue	Backbone	Detector	$mAP_{50:95}$(%)	mAP_{50}(%)	mAP_{75}(%)
FGFA* [45]	ICCV'17	ResNet-50	Faster R-CNN	47.1	74.7	52.0
SELSA* [40]	ICCV'19	ResNet-50	Faster R-CNN	48.6	78.4	52.5
MEGA* [4]	CVPR'20	ResNet-50	Faster R-CNN	48.1	77.3	52.2
TROI* [11]	AAAI'21	ResNet-50	Faster R-CNN	**48.8**	78.9	**52.8**
TransVOD [19]	ACM MM'21	ResNet-50	Deformable DETR	-	**79.9**	-
Frame Baseline [37]	CVPR'21	ResNet-50	Sparse R-CNN	48.7	71.1	52.4
SparseVOD	BMVC'22	ResNet-50	Sparse R-CNN	**54.7**	**80.3**	**60.1**
FGFA* [45]	ICCV'17	ResNet-101	Faster R-CNN	50.4	78.1	56.7
SELSA* [40]	ICCV'19	ResNet-101	Faster R-CNN	52.4	81.5	57.9
MEGA* [4]	CVPR'20	ResNet-101	Faster R-CNN	**53.1**	**82.9**	**59.1**
TROI* [11]	AAAI'21	ResNet-101	Faster R-CNN	51.6	82.6	56.3
MAMBA [36]	AAAI'21	ResNet-101	Faster R-CNN	-	**84.6**	-
TransVOD [19]	ACM MM'21	ResNet-101	Deformable DETR	-	81.9	-
Frame Baseline [37]	CVPR'21	ResNet-101	Sparse R-CNN	51.7	74.6	53.9
SparseVOD	BMVC'22	ResNet-101	Sparse R-CNN	**56.9**	81.9	**63.1**
FGFA* [45]	ICCV'17	ResNeXt-101	Faster R-CNN	52.5	79.6	59.8
SELSA* [40]	ICCV'19	ResNeXt-101	Faster R-CNN	**54.2**	83.1	**61.3**
MEGA [4]	CVPR'20	ResNeXt-101	Faster R-CNN	-	**84.1**	-
TROI* [11]	AAAI'21	ResNeXt-101	Faster R-CNN	54.4	**84.3**	60.9
Frame Baseline [37]	CVPR'21	ResNeXt-101	Sparse R-CNN	53.3	76.6	57.9
SparseVOD	BMVC'22	ResNeXt-101	Sparse R-CNN	**58.0**	83.1	**64.3**

Table 1: Comparison with other state-of-the-art methods on the ImageNet VID dataset. Results with * are reproduced. The two best results are highlighted in red and blue.

improvement on mAP @IoU=0.5, our SparseVOD demonstrates a significant increase (7.3 and 6 points) in the precise localization on mAPs @IoU=0.75 and @IOU=0.5:95 from the previous best method [11], reflecting the superiority of spatio-temporal learnable proposals.

When stronger backbones of ResNet-101 and ResNeXt-101 are incorporated into SparseVOD, the performance (mAP_{50}) further increases to 81.9% and 83.1%, respectively. Note that although MAMBA [36], MEGA [4], and TROI [11] demonstrate better results at mAP_{50}, our SparseVOD supersedes them by far (4∼5 points in mAP) on higher IoU thresholds. These results correspond to our argument that while prior VOD methods operating on dense to sparse detection pipelines show impressive results on mAP_{50}, they fail to produce confident and precise predictions. Furthermore, our SparseVOD boosts the single-frame baseline (Sparse R-CNN [37]) by a strong margin (8%∼9% mAP_{50}) with all backbone networks. This noticeable increase in mAP highlights the importance of leveraging temporal information to generate object proposals in VOD.

4.3 Iterative proposal Visualization Analysis

We visualize the behaviour of learned proposals boxes of a trained model on a video clip from the validation set in Fig. 3. Note that these proposals cover almost all potential regions in the target frames. This ensures high recall performance even with sparse proposals. Moreover, each stage in the cascaded architecture refines the bounding box offset and removes duplication. This makes our pipeline independent of any post-processing techniques to produce precise predictions. Fig. 3(d) further exhibits the robustness of our SparseVOD by producing high-quality predictions in challenging scenarios with camera defocus and part-occlusions. Please see supplementary materials for more qualitative analysis.

4.4 Ablation Studies

This section discusses the effect of key components and validates the design choices in our proposed method. Following [19], we perform all experiments on ImageNet VID dataset with ResNet-50 as the backbone. The run time (FPS) is tested on a single DGX A100 GPU. More ablation studies can be found in supplementary materials.

Effectiveness of each component in SparseVOD. Table 2 summarizes the impact of adding each component to build our proposed method. Beginning with the single-frame baseline, we incorporate the TFE module (explained in Section 3.1) and boost the AP_{50} from 71.1% to 76.9%. This demonstrates the benefit of exploiting inter-frame information to extract RoI proposal features. On the other hand, by separately plugging the SPFA module (explained in Section 3.1) into a single-frame baseline, we achieve substantial gains in AP_{50} from 71.1% to 79.1%. Finally, by combining both TFE and SPFA to build the proposed SparseVOD, we gain a further boost of 1.2% in AP_{50}, accomplishing 80.3%. These results establish the superiority of introducing spatio-temporal feature aggregation for learnable proposals.

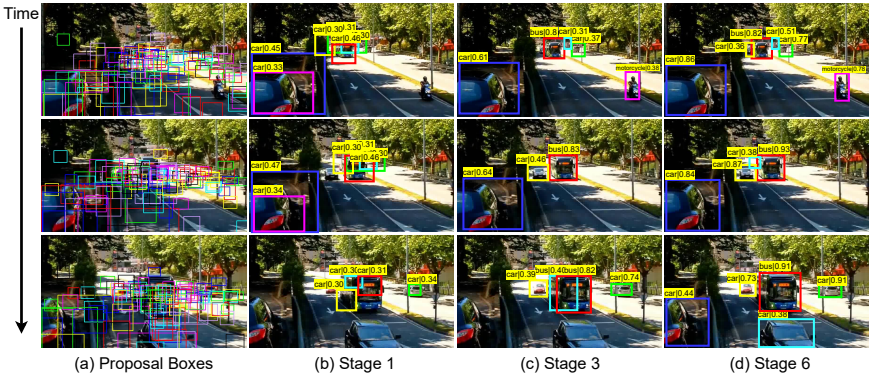

Figure 3: Illustration of the learned proposals and bounding box predictions at different iterative stages from a converged model. For brevity, we only visualize predictions with a confidence score greater than 0.3. Note that learned proposal boxes in (a) cover possible regions in all three video frames while the cascading heads in each stage enhance detections.

Single Frame Baseline	TFE	SPFA	AP_{50}(%)	AP_{75}(%)	$AP_{50:95}$(%)	FPS
✓	✗	✗	71.1	52.4	48.7	24.3
✓	✓	✗	$76.9_{\uparrow 5.8}$	$57.5_{\uparrow 5.1}$	$52.1_{\uparrow 3.4}$	14.9
✓	✗	✓	$79.1_{\uparrow 8.0}$	$59.0_{\uparrow 6.6}$	$54.1_{\uparrow 5.4}$	17.7
✓	✓	✓	$80.3_{\uparrow 9.2}$	$60.1_{\uparrow 7.7}$	$54.7_{\uparrow 6.0}$	14.4

Table 2: Ablation on effectiveness of each module in SparseVOD.

Stages	AP_{50}(%)	AP_{75}(%)	FPS
1	60.0	27.2	42.2
2	74.5	52.8	29.4
3	78.0	58.3	23.7
4	78.5	58.6	19.8
5	79.1	59.8	16.1
6	**80.3**	**60.1**	**14.4**
12	77.1	55.2	5.7

Table 3: Ablation on the number of stages.

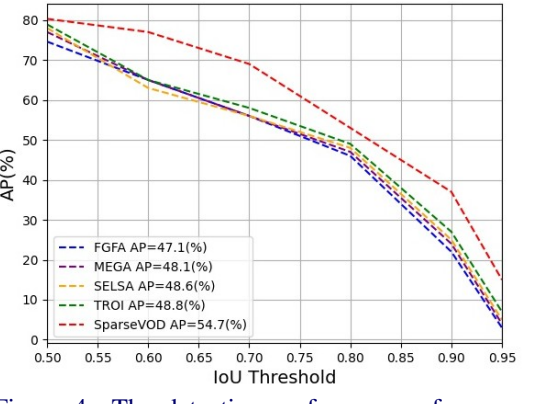

Figure 4: The detection performance of recent VOD methods and our SparseVOD on increasing IoU threshold.

TFE (Avg.)	TFE (MSA)	SPFA (Avg.)	SPFA (MSA)	AP_{50}(%)	AP_{75}(%)	$AP_{50:95}$(%)	FPS
✗	✗	✗	✗	71.1	52.4	48.7	24.3
✓	✗	✓	✗	75.8	52.9	49.1	18.5
✓	✗	✗	✓	79.4	58.9	54.1	16.3
✗	✓	✓	✗	78.5	58.1	52.1	14.9
✗	✓	✗	✓	**80.3**	**60.1**	**54.7**	14.4

Table 4: Ablation on the effectiveness of multi-head attention in TFE (Temporal Feature Extraction) and SPFA (Semantic Proposal Feature Aggregation) modules. The first row represents results from a single-frame baseline. The terms Avg. and MSA denote simple averaging and multi-head self-attention.

N_{ref}	AP_{50}(%)	AP_{75}(%)	FPS
1	71.1	52.4	24.3
2	79.0	58.2	17.7
4	79.9	59.5	15.9
6	**80.3**	**60.1**	**14.4**
10	80.2	60.2	11.7
14	80.3	60.2	9.5

Table 5: Ablation on number of support frames.

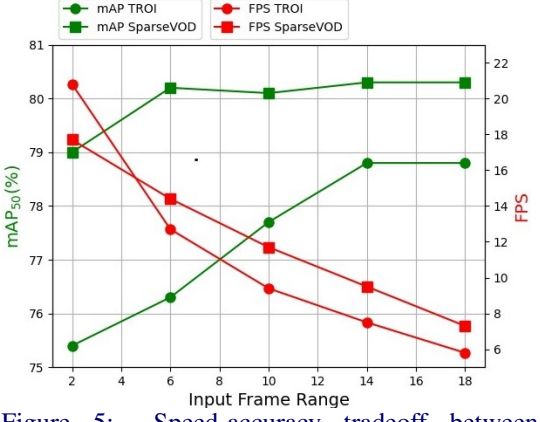

Figure 5: Speed-accuracy tradeoff between SparseVOD and previous best competitor (TROI+SELSA [11]).

Number of Stages. The impact of increasing stages in an iterative architecture is summarized in Table 3. Note that without iterative architecture, even though the performance AP_{50}(%) reaches 60.0, the AP_{75} is merely 27.2%. Since the input proposal boxes at the first stage are just random distribution of possible object locations, this result (AP_{50}=60.0%) indicates that computing AP on a single IoU threshold of 0.5 is not a reliable evaluation metric. By increasing iterative stages to 3, the performance already reaches a comparable AP_{50} of 78% and surpasses prior methods on AP_{75} with 58.3%. Finally, similar to [37], the performance saturates at 6 stages. Hence, we adopt 6 stages in our experiments.

Comparing High-Quality Detection. Fig. 4 shows the AP curves of recent state-of-the-art VOD methods and our SparseVOD under increasing IoU thresholds. It is evident that the proposed method consistently outperforms prior works with a significant margin on all the evaluation metrics. Note that although the difference is mild with the previous best competitor TROI [11], on low IoU threshold (0.5), it rises on higher IoU thresholds. These results reflect the superiority of sparse spatio-temporal learnable proposals over hand-crafted dense priors optimized on a single IoU level in existing VOD methods.

Effectiveness of Multi-head Attention in TFE and SPFA. We demonstrate the effectiveness of multi-head attentional blocks [38] in Temporal Feature Extraction (TFE) and Semantic Proposal Feature Aggregation (SPFA) modules in Table 4. For direct comparison, we conduct baseline experiments where attentional weights in TFE and SPFA are replaced by simple averaging to aggregate features from target and support frames. As shown in Ta-

ble 4, with averaging in both modules, the AP_{50} reaches 75.8%, reflecting the benefit of leveraging temporal information to refine object proposals. Furthermore, we conduct experiments by switching attention to averaging in one of the two modules. With TFE (Avg) and SPFA (MSA), we observe an impressive speed-accuracy tradeoff of 79.4% AP_{50} and 16.3 FPS. Since the averaging in TFE is performed on the most similar RoI features computed with cosine similarities, when combined with SPFA (MSA), it already provides an acceptable object feature representation. However, these results are still inferior (-0.9 points in AP_{50}) to the performance achieved when multi-head attention is plugged in both TFE and SPFA. These results (80.3% AP_{50}) indicate the superiority of multi-head attentional aggregation in our method.

Number of Support Frames. Table 5 presents the ablations on the number of support frames. We follow the identical frame sampling strategy as in [11, 42], where support frames are uniformly sampled from the entire video. We can see that the AP_{50} already reaches 79.0% with 2 support frames. Upon increasing support frames, the performance keeps increasing and tends to stabilize after reaching the AP_{50} of 80.3% with 6 support frames.

Speed-accuracy Tradeoff. Table 2 shows that the computational load in our SparseVOD stems from Temporal Feature Extraction (TFE) and Semantic Proposal Feature Aggregation (SPFA). Since the results of [19] are not reproducible, for direct comparison, we analyse the speed-accuracy tradeoff of the second best method TROI [11] and our SparseVOD in Fig. 5. Note that TROI [11] is built upon SELSA [42] to enhance performance. With only 6 support frames sampled from the entire video, our SparseVOD achieves a new best AP_{50} of 80.3% with a run time of 14.4 FPS. In contrast, TROI manages to reach its best performance (AP_{50} of 78.8%) with a run time of 7.5 FPS after utilizing 14 support frames. The results in Fig. 5 demonstrate that the temporal feature aggregation from several support frames in prior works [11, 42] lead to a major increase in run time, producing a sub-optimal speed-accuracy tradeoff. Contrarily, thanks to the spatio-temporal learnable proposals, our SparseVOD yields an optimal speed-accuracy tradeoff (79.0% AP_{50} and 17.7 FPS) with merely 2 support frames.

5 Conclusion

This paper proposes SparseVOD, a novel video object detection pipeline which introduces spatio-temporal feature aggregation to refine object proposals. The SparseVOD effectively eliminates hand-crafted dense priors and provides reliable proposal features even with deteriorated input frames. Particularly, the SparseVOD incorporates attention-guided Temporal Feature Extraction and Semantic Proposal Feature Aggregation modules in Sparse R-CNN [37]. Extensive experiments validate that SparseVOD significantly improves the baseline performance by 8%-9% in mAP and achieves the state-of-the-art 80.3% mAP_{50} on the ImageNet VID dataset with ResNet-50 backbone. Besides, our SparseVOD beats existing methods in terms of high-quality predictions and optimal speed-accuracy tradeoff. To our knowledge, our work is the first one that exploits temporal information in directly generating a sparse set of object proposals for video object detection. We hope similar work can be applied to other video analysis tasks like object tracking and video instance segmentation.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern*

recognition, pages 6154–6162, 2018.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Kai Chen, Jiaqi Wang, Shuo Yang, Xingcheng Zhang, Yuanjun Xiong, Chen Change Loy, and Dahua Lin. Optimizing video object detection via a scale-time lattice. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7814–7823, 2018.

[4] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10337–10346, 2020.

[5] Yiming Cui, Liqi Yan, Zhiwen Cao, and Dongfang Liu. Tf-blender: Temporal feature blender for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8138–8147, 2021.

[6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016.

[7] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Object guided external memory network for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6678–6687, 2019.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[10] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE international conference on computer vision*, pages 3038–3046, 2017.

[11] Tao Gong, Kai Chen, Xinjiang Wang, Qi Chu, Feng Zhu, Dahua Lin, Nenghai Yu, and Huamin Feng. Temporal roi align for video object recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1442–1450, 2021.

[12] Chaoxu Guo, Bin Fan, Jie Gu, Qian Zhang, Shiming Xiang, Veronique Prinet, and Chunhong Pan. Progressive sparse local attention for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3909–3918, 2019.

[13] Liang Han, Pichao Wang, Zhaozheng Yin, Fan Wang, and Hao Li. Exploiting better feature aggregation for video object detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1469–1477, 2020.

[14] Liang Han, Pichao Wang, Zhaozheng Yin, Fan Wang, and Hao Li. Class-aware feature aggregation network for video object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[15] Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. Mining inter-video proposal relations for video object detection. In *European conference on computer vision*, pages 431–446. Springer, 2020.

[16] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465*, 2016.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[19] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. End-to-end video object detection with spatial-temporal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1507–1516, 2021.

[20] Zhengkai Jiang, Peng Gao, Chaoxu Guo, Qian Zhang, Shiming Xiang, and Chunhong Pan. Video object detection with locally-weighted deformable neighbors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8529–8536, 2019.

[21] Zhengkai Jiang, Yu Liu, Ceyuan Yang, Jihao Liu, Peng Gao, Qian Zhang, Shiming Xiang, and Chunhong Pan. Learning where to focus for efficient video object detection. In *European conference on computer vision*, pages 18–34. Springer, 2020.

[22] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 817–825, 2016.

[23] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. Object detection in videos with tubelet proposal networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 727–735, 2017.

[24] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2896–2907, 2017.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[30] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[32] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3):211–252, 2015.

[34] Alberto Sabater, Luis Montesano, and Ana C Murillo. Robust and efficient post-processing for video object detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10536–10542. IEEE, 2020.

[35] Mykhailo Shvets, Wei Liu, and Alexander C Berg. Leveraging long-range temporal relationships between proposals for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9756–9764, 2019.

[36] Guanxiong Sun, Yang Hua, Guosheng Hu, and Neil Robertson. Mamba: Multi-level aggregation via memory bank for video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2620–2627, 2021.

[37] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[39] Thang Vu, Hyunjun Jang, Trung X Pham, and Chang Yoo. Cascade rpn: Delving into high-quality region proposal network with adaptive convolution. *Advances in neural information processing systems*, 32, 2019.

[40] Chi Wang, Yang Hua, Zheng Lu, Jian Gao, and Neil Robertson. Temporal meta-adaptor for video object detection. In *British Machine Vision Conference 2021*, 2021.

[41] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 542–557, 2018.

[42] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9217–9225, 2019.

[43] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[44] Tong Yang, Xiangyu Zhang, Zeming Li, Wenqiang Zhang, and Jian Sun. Metaanchor: Learning to detect objects with customized anchors. *Advances in neural information processing systems*, 31, 2018.

[45] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 408–417, 2017.

[46] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2349–2358, 2017.

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.