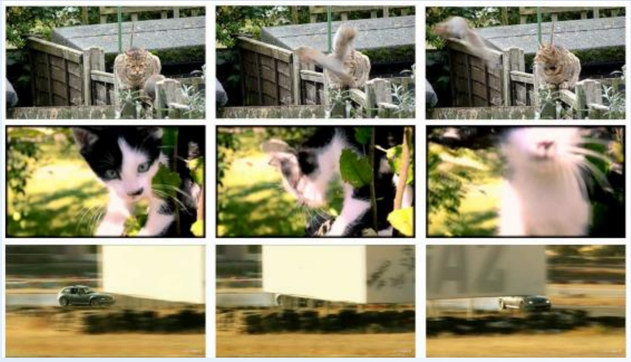
Spatio-Temporal Learnable Proposals for End-to-End Video Object Detection

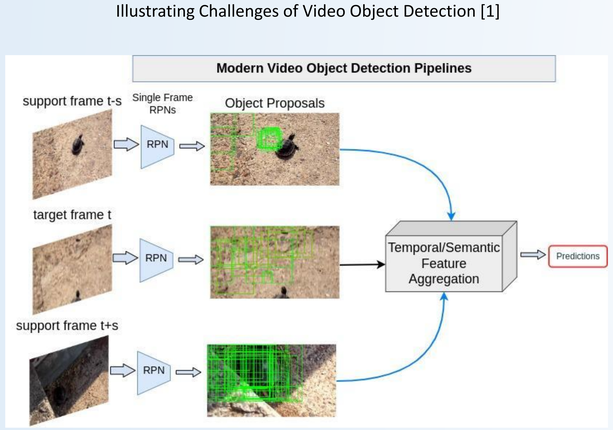
Khurram Azeem Hashmi Didier Stricker Muhammad Zeshan Afzal

German Research Center for Artificial Intelligence (DFKI)

Motivation



Illustrating Challenges of Video Object Detection [1]

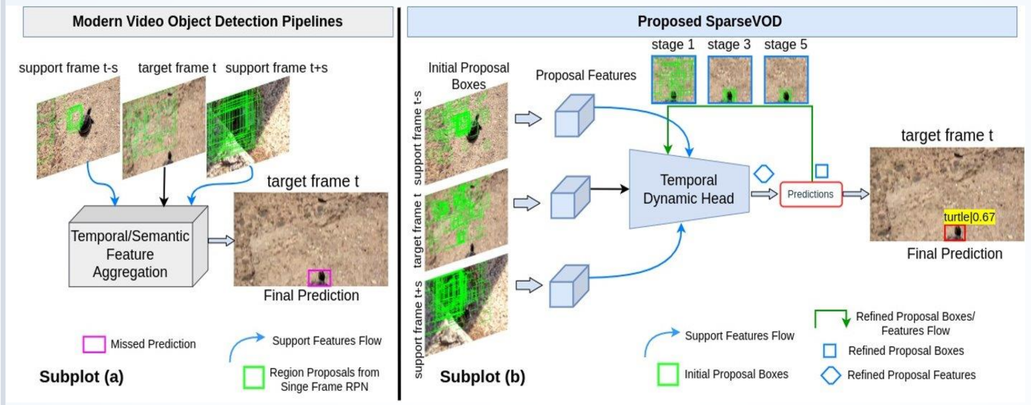


Prior works [1,2,3,4] mainly enhance the target frame feature representation by aggregating features from support frames.

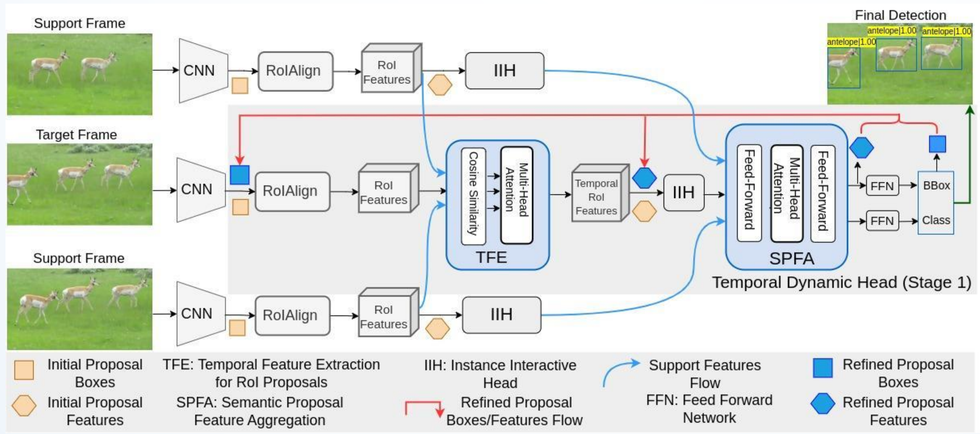
Contributions

- **SparseVOD, a novel end-to-end trainable** video object detection method.
- We propose an idea of **exploiting temporal information to learn region proposals.**
- **We exceed state-of-the-art** methods on increasing IoU thresholds (IoU > 0.5) and achieve **optimal speed-accuracy tradeoff.**

Prior Works Against Our Method



Our Proposed SparseVOD



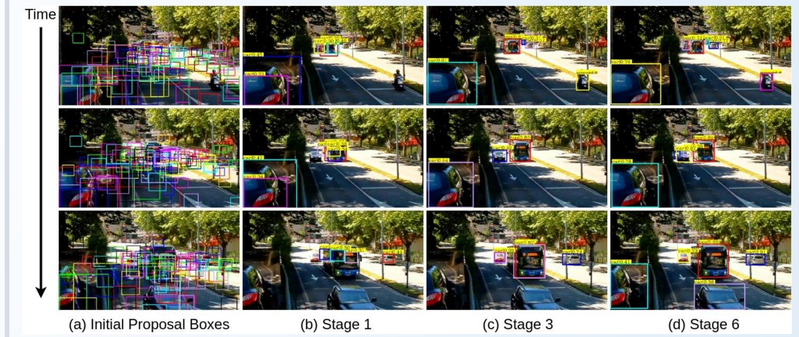
Effectiveness of TFE and SPFA

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

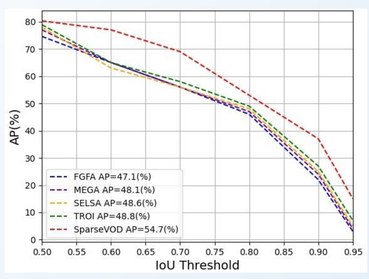
Comparison with Previous Methods on ImageNet VID

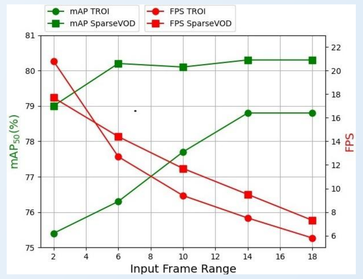
Methods	Venue	Backbone	Detector	$mAP_{50:95}(\%)$	$mAP_{50}(\%)$	$mAP_{75}(\%)$
FGFA* [45]	ICCV'17	ResNet-50	Faster R-CNN	47.1	74.7	52.0
SELSA* [42]	ICCV'19	ResNet-50	Faster R-CNN	48.6	78.4	52.5
MEGA* [4]	CVPR'20	ResNet-50	Faster R-CNN	48.1	77.3	52.2
TROI* [11]	AAAI'21	ResNet-50	Faster R-CNN	**48.8**	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* [42]	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* [42]	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**

Behaviour of Learned Proposals



High Quality Detections and Speed-Accuracy Tradeoff





References

[1] Wu, H., Chen, Y., Wang, N., & Zhang, Z. (2019). Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9217-9225).

[2] Gong, T., Chen, K., Wang, X., Chu, Q., Zhu, F., Lin, D., ... & Feng, H. (2021, May). Temporal ROI align for video object recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 2, pp. 1442-1450).

[3] Han, L., Wang, P., Yin, Z., Wang, F., & Li, H. (2021). Class-aware feature aggregation network for video object detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

[4] Han, M., Wang, Y., Chang, X., & Qiao, Y. (2020, August). Mining inter-video proposal relations for video object detection. In *European conference on computer vision* (pp. 431-446). Springer, Cham.