Video Object Detection: A Pilot Study

Abhishek Rajora
Indian Institute of Technology, Jodhpur
rajora.1@iitj.ac.in

Abu Shahid
Indian Institute of Technology, Jodhpur
shahid.3@iitj.ac.in

Abstract

This paper aims to explore various state-of-the-art object detection models for video analysis, including Mask R-CNN[3], Temporal ROI Align[7], RetinaNet[2], Boxmask[8], YOLOV[4], TransVOD[5] and VSTAM[6]. The paper provides a comprehensive analysis of these models, studying their architectures, design choices, and performance on the IILVSRC2016-VID[1] dataset. The dataset is described in detail, including its specifications, which cover common objects/subjects of 35 categories and predicates of 132 categories. The paper evaluates the models quantitatively and discusses the results obtained. Additionally, the paper establishes relationships between these architectures, highlighting their similarities and differences, and providing insights into their strengths and weaknesses. Specifically we focused on integrating box masks in the state-of-the-art models to achieve even higher mAP scores. Overall, this paper serves as a valuable resource for researchers and practitioners working on video object detection and analysis, providing a comprehensive overview of the latest techniques and approaches in the field.

1. INTRODUCTION

Object detection in videos has been a challenging task in computer vision due to the complex nature of video data, which contains a large number of objects, different scales, orientations, occlusions, and background variations. Therefore, it requires the development of robust and efficient algorithms that can handle such variations and provide accurate and timely object detection and tracking. One of the key factors that have enabled this progress is the use of backbone architectures, which are pre-trained deep neural networks that can extract high-level features from images and videos. Backbone architectures have become an essential component of many state-of-the-art video object detection models, as they can provide a good trade-off between accuracy and speed.

Among the commonly used backbones in video object detection, Fast R-CNN and FCN are notable examples. Fast R-CNN is a meta-algorithm that builds on the earlier R-CNN and Faster R-CNN models, by introducing a region proposal network (RPN) that can generate object proposals directly from feature maps. Fast R-CNN is known for its intuitive design and good speed, making it a popular choice for many video object detection applications. These methods are conceptually intuitive and offer flexibility and robustness, together with fast training and inference time. Several state-of-the-art models have been proposed that utilizes these backbones and achieve high accuracy and real-time performance for video object detection, such as Mask R-CNN, Temporal ROI Align, BoxMask, YOLOV, TransVOD and VSTAM. Our goal in this work is to study these models and provides insights in comparably enabling framework for object detection.

In this paper, we aim to provide a comprehensive overview of these models and evaluate their performance on the IILVSRC2016-VID dataset, which is a widely used benchmark for video object detection. The IILVSRC2016-VID dataset consists of 1000 video sequences of 35 categories of common objects/subjects and 132 categories of predicates. The dataset is split into 800 training sets and 200 test sets, making it suitable for evaluating the performance of different models.

The paper is organized as follows: In section 2, we discuss the related work on object detection and tracking in videos. In section 3, we provide a detailed description of the models and techniques used in this paper. In section 4, we present the experimental setup and results of our evaluation on the IILVSRC2016-VID dataset. In section 5, we discuss the results obtained and provide insights into the strengths and weaknesses of the different models. Finally, in section 6, we conclude the paper and discuss future directions for research in this area.

2. Related Work

Object detection and tracking in videos have been extensively studied in the literature, with several approaches proposed over the years. Early approaches used handcrafted features and background subtraction techniques to detect and track objects in videos. However, these approaches suffered from low accuracy and were limited to simple scenarios.

Viola-Jones. The initial approach for video object detection was based on still images, and relied on handcrafted features and heuristics. The Viola-Jones detector, proposed in 2001, was a landmark work in this field, as it introduced a simple and efficient algorithm for object detection for still images using Haar features and Adaboost classifier. The Viola-Jones detector achieved high accuracy and speed, and was widely used in many applications, such as face recognition and video surveillance. However, the Viola-Jones detector was limited to detecting faces and other simple objects, and its performance degraded in the presence of occlusions, cluttered backgrounds, and variations in object pose and scale. Therefore, there was a need for more powerful and flexible models that can handle these challenges and scale to large datasets.

One-stage Detectors. One-stage detectors, such as YOLO (You Only Look Once) and SSD (Single Shot Detector), are designed to detect objects in a single pass through the network, by predicting object bounding boxes and class labels directly from feature maps. These detectors use a grid-based approach to generate object detections directly from the feature maps, without the need for RoI operations. In the grid-based approach, the input frames are divided into a set of fixed-size grids, and each grid is associated with a set of anchor boxes of different aspect ratios and scales. The object detections are generated by predicting the class probabilities and bounding box offsets for each anchor box, and then selecting the anchor boxes with the highest scores as the final object detections. One-stage detectors have a simpler and faster design compared to two-stage detectors, and can achieve real-time performance on high-resolution video streams. However, one-stage detectors may suffer from lower accuracy and precision, especially for small objects and complex scenes.

Two-stage Detectors. Two-stage detectors, such as Faster R-CNN (Region-based Convolutional Neural Network) and Mask R-CNN, are based on a region proposal network (RPN) that generates object proposals from feature maps, followed by a classifier that predicts the object class and refine the bounding box. In the first step, a region proposal network (RPN) is used to generate a set of candidate regions (or proposals) that may contain objects. The RPN typically uses a sliding window approach to generate proposals at different scales and aspect ratios. In the second step, a separate neural network is used to classify the proposals and refine their bounding boxes. Two-stage detectors are generally more accurate than one-stage detectors, but are also slower and more computationally expensive.

Feature Maps. Feature maps are an important component in many computer vision tasks, including object detection and recognition. In video object detection, feature maps can be used in combination with Region of Interest (RoI) operations to extract features and classify objects within specific regions of the input frames. RoI operations involve defining a bounding box around a specific object or region of interest in the input frames, and extracting the corresponding features from the feature maps within that bounding box. This is done by mapping the coordinates of the bounding box onto the feature maps and extracting the features within that region. RoI operations is used for two-stage object detection models.

Transformers. In recent advancement, transformer-based frameworks have been proposed for video object detection, such as TransVOD and VSTAM. These frameworks use attention guided mechanisms to capture long-term dependencies and object motion and appearance changes over time, achieving state-of-the-art results on several benchmarks.

3. Literature Review

Our literature review extensively references a wide range of papers to provide comprehensive insights and understandings of the underlying architecture in them.

3.1. Mask R-CNN

Mask R-CNN is based on the Faster R-CNN architecture, which consists of two main components: a region proposal network (RPN) and a fast R-CNN detector. The RPN generates candidate object proposals, and the fast R-CNN detector uses these proposals to predict the class labels and bounding boxes for the objects. Mask R-CNN extends this architecture by adding a third branch to the network, which performs instance segmentation on the object proposals. The instance segmentation branch in Mask R-CNN is based on the fully convolutional network (FCN) architecture, which was originally developed for semantic segmentation tasks. The FCN consists of a series of convolutional and deconvolutional layers that produce a dense pixel-wise prediction for each image.

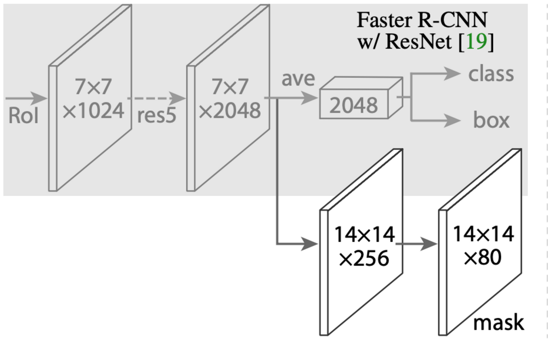

Figure 1. Mask R-CNN Architecture

From Figure.1, we can see that FCN is modified to take the RoI features from the object proposals generated by the RPN, and to produce a binary mask for each object instance. The backbone of Mask R-CNN is based on the ResNet architecture, which consists of a series of residual blocks that allow for deeper and more efficient network training. Additionally, it incorporates RoIAlign instead of RoIPooling for more accurate feature extraction from the object proposals to preserve spatial information, and a mask head network for generating the instance segmentation masks.

3.2. Temporal ROI Align

It extracts the most similar ROI features from support frames for target frame proposals based on feature similarity, implicitly incorporating temporal information. A temporal attention mechanism is employed to aggregate these ROI features, giving more importance to clear object instances. The proposed operator not only improves performance in video object detection but also has potential applications in other video tasks, such as video instance segmentation.

3.3. RetinaNet

The RetinaNet architecture is based on a feature pyramid network (FPN) that extracts features from images at multiple scales. The FPN is combined with a single-stage detection architecture, which allows for a simpler and more efficient model compared to previous two-stage detection models like Faster R-CNN. The RetinaNet model uses a novel focal loss function to address the class imbalance problem in dense object detection, where the vast majority of regions in an image do not contain any objects.

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t) \tag{1}$$

, where p_t is the predicted probability of the ground-truth class, and γ is a tunable parameter that controls the focusing factor. The focal loss function assigns a higher weight to misclassified examples with low predicted probabilities, which helps the model to better learn from hard examples.

3.4. BoxMask

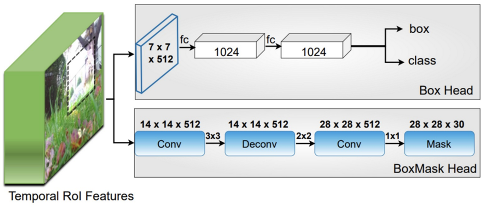

Figure 2. Overall architecture of the detection phase equipped with a BoxMask head at the bottom.. Due to its simplistic design, the proposed BoxMask head can be integrated into any region-based VOD method.

The BoxMask model builds upon previous work in object detection, incorporating key ideas from the two-stage detection model Mask R-CNN and the single-stage detector RetinaNet. Specifically, BoxMask uses a two-stage detection framework that consists of a region proposal network (RPN) and a box and mask head network. The RPN generates candidate object regions in each frame of the video, while the box and mask head network classifies and refines these regions to produce the final object detections.

One key innovation of the BoxMask model is its use of bounding box supervision, which involves explicitly modeling the relationships between object boxes across frames. This approach allows the model to better handle complex motion patterns and occlusions, resulting in more accurate and robust object detections in videos. The BoxMask model achieves this by incorporating a novel box and mask consistency loss, which encourages the predicted object boxes and masks to be consistent across frames.

The BoxMask model employs a bounding box mask tensor Mbox to optimize the mask prediction by minimizing the cross-entropy loss Lbm, which is defined by the following equation:

$$L_{bm} = -\frac{1}{m} \sum_{i=1}^{m} \sum_{c=0}^{C-1} M(i,c) \log(y(i,c)) \tag{2}$$

Here, L_{bm} is the loss function for predicting the class for each pixel in each sampled Region of Interest (RoI). The term C represents the number of classes, and y(i, c) denotes the predicted probability of class c for pixel i. The BoxMask loss function decouples the prediction of mask and class labels, allowing the model to learn features for localization.

The BoxMask model integrates the BoxMask head in region-based video object detection methods to compute the detection loss L_{det}, which is defined as follows:

$$L_{det} = L_{cls} + L_{reg} + \lambda L_{bm} \tag{3}$$

Here, L_{cls} and L_{reg} are the classification and regression losses, respectively.

3.5. YOLOV

Traditional two-stage approaches suffer from slow speed due to the large number of low-confidence region candidates. To overcome this limitation, the authors introduce a two-stage pipeline where the first stage involves prediction and discarding regions with low confidences, while the second stage focuses on region-level refinement through temporal aggregation.

The authors emphasize the importance of seeking supportive information from other frames for a target frame (keyframe) in video object detection. By designing their approach as a region/feature selection after the prediction head of one-stage detectors, they aim to benefit from the efficiency of one-stage detectors and the accuracy gained from temporal aggregation. The proposed strategy can be applied to various base detectors such as FCOS and PPY-OLOE.

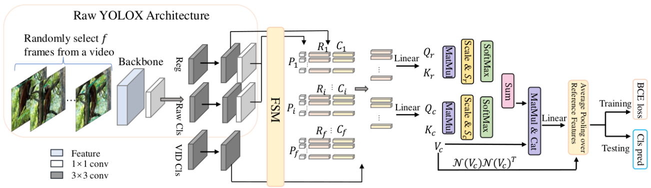

Figure 3. YOLOV Architecture

A Feature Selection Module (FSM) is introduced to select high-quality candidates from the feature maps of the detection head. This module picks top predictions based on confidence scores, applies Non-Maximum Suppression (NMS) to reduce redundancy, and collects the features of these predictions for further refinement.

3.6. TransVOD

The paper introduces TransVOD, an end-to-end video object detection system that utilizes spatial-temporal Transformer architectures. The main objective of the paper is to streamline the pipeline of video object detection by eliminating the need for hand-designed components used for feature aggregation, such as optical flow models and relation networks. TransVOD leverages the object query design from DETR, which eliminates the need for post-processing methods like Seq-NMS.

The proposed approach incorporates a temporal Transformer consisting of two components: the Temporal Query Encoder (TQE) and the Temporal Deformable Transformer Decoder (TDTD). The TQE fuses object queries, while the TDTD is responsible for obtaining the detection results for each frame. These design choices significantly improve the performance of the baseline deformable DETR by achieving a notable increase in mean Average Precision (mAP) of 3%-4% on the ImageNet VID dataset.

3.7. VSTAM

The VSTAM model employs a sparse attention mechanism that selectively attends to relevant frames and features in the input video. This attention mechanism is guided by the memory module, which consists of a set of learnable keys and values that encode information about the objects and their trajectories. The attention mechanism is used to compute a set of weights that determine the relevance of each frame and feature for the task at hand. This enables the model to focus its attention on the most relevant frames and features, reducing the computational complexity of processing long video sequences. It also employs random and positional attention mechanisms, which enable the model to learn spatial and temporal dependencies in the input video.

4. Experimental Setup

To evaluate the performance of various video object detection models, we utilized a pytorch Dataset that gets video data from the ImageNet Video dataset[1]. The dataset contains 1,000 video sequences, split into 800 for training and 200 for testing, and covers 35 object categories and 132 predicate categories.

We performed necessary preprocessing on the dataset, such as resizing the frames to 224x224 sized pixels and splitting the videos into individual frames. We also streamlined our workflow by creating a Dataloader that feeds the data into the models for training and testing.

For all the models, we implemented the experimental setup using hyperparameters provided in their respective official papers. This includes the use of specific network architectures such as ResNet, ResNeXt, and FPN, as well as the size of input images and any changes made to the focal loss. We also used other methods described in the papers such as the use of a size window in TransVOD and VSTAM. These hyperparameters were carefully chosen to ensure that our experiments were consistent with those reported in the literature and to provide a fair comparison between the models.

For evaluation metric, we are using Intersection over Union (IoU) with a threshold of 0.5 as the evaluation metric for all the models. This is a commonly used metric for object detection tasks that measures the overlap between the predicted bounding box and the ground truth bounding box. We also use mean Average Precision (mAP) as the secondary evaluation metric. mAP is a widely used metric that takes into account both precision and recall of object detection models. It measures the accuracy of the model across multiple IoU thresholds, typically ranging from 0.5 to 0.95.

With the basic workflow ready and the baseline performance established, we can now proceed to implement the models described in the literature review and compare their performance against the baseline.

5. Comparative Study and Analysis

Quantitative Results. The ImageNet VID dataset is one of the standard benchmarks for object detection, and several state-of-the-art methods have been developed to tackle this task. The mAP (R-50) and $\text{IoU}_0.5$ scores are the commonly used metrics to evaluate the performance of object detection models on this dataset.

Among the compared methods, VSTAM achieves the highest mAP score of 91.1, followed closely by TransVOD at 90.0. Both methods use visual transformers, which have shown promising results in many computer vision tasks, including object detection. These results suggest that visual transformers are a powerful tool for object detection, and they outperform the traditional backbone networks like ResNet and ResNext used in other methods.

The next best performing method is TROI, which achieves an mAP score of 78.9, followed by TROI + BoxMask at 80.7. These methods use temporal RoI align, which takes advantage of the temporal consistency of object motion in video data to improve object detection performance. RetinaNet and RetinaNet + BoxMask achieve mAP scores of 61.1 and 62.7, respectively.

The two versions of YOLOV (YOLOV and TinyYOLOV) achieve the lowest mAP scores among the compared methods, at 54.9 and 51.4, respectively. However, when combined with the BoxMask approach, the performance of YOLOV and TinyYOLOV is improved, but still, they remain the lowest performing methods among the compared ones.

Regarding $\text{IoU}_{0.5}$ scores, VSTAM and TransVOD achieve the highest scores of 0.83 and 0.85, respectively. The $\text{IoU}_{0.5}$ scores for other methods are lower, ranging from 0.52 for TinyYOLOV + BoxMask to 0.81 for TROI.

TABLE I. Comparison of existing state-of-art methods on ImageNet VID Dataset.

Method	**mAP (R-50)**	**IoU$_{0.5}$**
Mask R-CNN	59.5	0.6
TROI	78.9	0.81
RetinaNet	61.1	0.67
YOLOV	54.9	0.63
TinyYOLOV	51.4	0.55
TransVOD	90.0	0.85
VSTAM	91.1	0.83
TROI + BoxMask	80.7	0.74
RetinaNet + BoxMask	62.7	0.69
YOLOV + BoxMask	57.2	0.59
TinyYOLOV + BoxMask	53.2	0.52

Table 2. shows the results of experiments performed on RetNet architecture with focal loss. The experiments were performed by varying the parameters α and γ while keeping the value of IoU threshold constant at 0.5. The mAP (R-50) was used as the evaluation metric for the experiments.

From the results, we can observe that as the value of α increases, the mAP (R-50) also increases. This is expected as increasing the value of α puts more emphasis on hard examples, leading to better classification of such examples. The highest mAP (R-50) of 52.5 is obtained when α is set to 2.

However, when γ is set to 0.5, the mAP (R-50) drops compared to when γ is set to 0.25. This is expected as a higher value of γ is expected to increase the weight given to positive samples, leading to better classification of such samples but keeping it much high will cause negative samples to diminish and thus there will be false positives. The highest mAP (R-50) of 52.5 is obtained when γ is set to 0.25.

We can also observe that when α is set to 5, the mAP (R-50) drops significantly compared to the other values of α. This may be due to too much emphasis on hard examples, causing the model to overfit on the training set and perform poorly on the test set.

TABLE II. mAP analysis by varying α and γ weights for focal loss in RetinaNet architecture.

α	γ	**AP (R-50)**
0	0.75	49.4
0.1	0.75	49.9
0.2	0.75	50.7
0.5	0.5	51.7
1	0.25	52.0
2	0.25	52.5
5	0.25	49.6

Qualitative Analysis. The BoxMask + RetinaNet model has been successful in detecting objects with good bounded boxes as can be seen in Figure 4. But we can also observe that there are still some areas of improvement for this model. In situations where there is occlusion or poor lighting, the model can sometimes struggle to accurately detect objects.

Let us consider an example to illustrate the performance of the three object detection models. In Figure 5, 6, and 7, the video frame contains two persons, two bikes, and a backpack. We tested the models' performance on this frame, and observed that TinyYOLOv3 had the fastest inference time but detected an extra motorbike, which was not present in the frame.

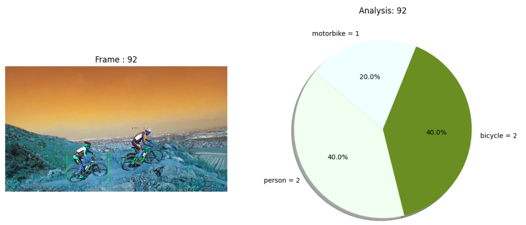

Figure 5. TinyYOLOv3

YOLOv3, on the other hand, correctly identified the persons and bikes, but failed to detect the backpack.

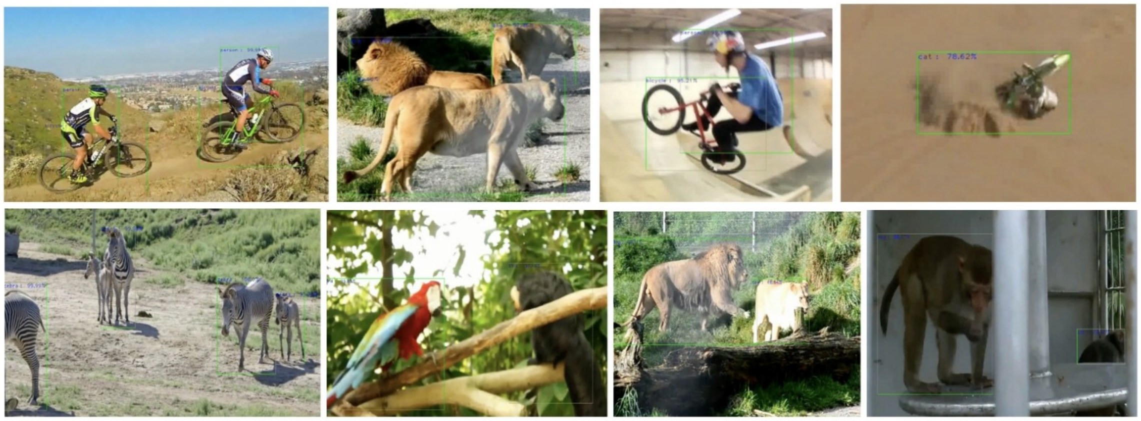

Figure 4. Bounded Boxes for detected objects from BoxMask + RetinaNet architecture

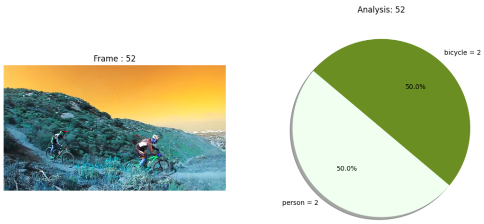

Figure 6. YOLOV

Finally, RetinaNet was able to detect all five objects correctly, indicating superior performance compared to the other models.

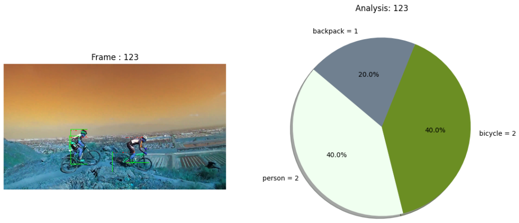

Figure 7. RetinaNet

6. Conclusion

In conclusion, we have made significant progress towards achieving our goal of exploring various state-of-the-art object detection models for video analysis. We have developed a dataloader that streamlines our workflow and enables us to efficiently process data from the IILVSRC2016-VID dataset. The study also explored the impact of BoxMasking on the performance of these models. The results showed that BoxMasking can improve the performance of RetinaNet and YOLOV. Furthermore, the BoxMasking approach also has the potential to improve the performance of other object detection models. However, the results also indicated that BoxMasking may fail in occlusion or lighting conditions, and further improvement is needed to increase its efficiency.

Overall, the study suggests that BoxMasking can be a promising approach to improve object detection models' performance. Furthermore, the study highlights the importance of selecting appropriate object detection models based on the specific use case, data characteristics, and evaluation metrics. Future research can explore more in-depth analysis of BoxMasking and other approaches to improve object detection models' performance.

Insights can be implemented using BoxMask include, but not limited to, improving object detection models' performance, reducing false positives and false negatives, and enhancing object localization accuracy using frame window. Additionally, BoxMasking can also help in better detecting and segmenting objects in challenging conditions, such as low lighting and occlusion, leading to better object detection and tracking in real-world applications.

References

[1] Hao Su Jonathan Krause Sanjeev Satheesh Sean Ma Zhiheng Huang Andrej Karpathy Aditya Khosla Michael Bernstein Alexander C. Berg Li Fei-Fei Olga Russakovsky, Jia Deng. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision volume, IJCV*, 2015.

[2] Ross Girshick Kaiming He Piotr Dollar Facebook AI Research (FAIR) Tsung-Yi Lin, Priya Goyal. Focal loss for dense object detection. *arXiv*, 2018.

[3] Piotr Dollar Ross Girshick Facebook AI Research (FAIR) Kaiming He, Georgia Gkioxari. Mask r-cnn. *arXiv*, 2018.

[4] Xiaojie Guo1 Yuheng Shi1, Naiyan Wang2. Yolov: Making still image object detectors great at video object detection. *arXiv*, 2023.

[5] Lu He Yibo Yang Guangliang Cheng Yunhai Tong Lizhuang Ma† Dacheng Tao Fellow Qianyu Zhou, Xiangtai Li. Transvod: End-to-end video object detection with spatial-temporal transformers. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2022.

[6] AKIHIRO SUGIMOTO MASATO FUJITAKE. Video sparse transformer with attention-guided memory for video object detection. *IEEE Access*, 2022.

[7] Xinjiang Wang Qi Chu Feng Zhu Dahua Lin Nenghai Yu Huamin Feng Tao Gong, Kai Chen. Temporal roi align for video object recognition. *arXiv*, 2021.

[8] Didier Stricker Muhammamd Zeshan Afzal Khurram Azeem Hashmi, Alain Pagani. Boxmask: Revisiting bounding box supervision for video object detection. *arXiv*, 2022.

[1] [2] [3] [4] [5] [6] [7] [8]