

BoxMask: Revisiting Bounding Box Supervision for Video Object Detection

Khurram Azeem Hashmi Alain Pagani Didier Stricker
Muhamamd Zeshan Afzal

DFKI - German Research Center for Artificial Intelligence

firstname[0]_firstname[1].lastname@dfki.de

Abstract

We present a new, simple yet effective approach to uplift video object detection. We observe that prior works operate on instance-level feature aggregation that imminently neglects the refined pixel-level representation, resulting in confusion among objects sharing similar appearance or motion characteristics. To address this limitation, we propose BoxMask, which effectively learns discriminative representations by incorporating class-aware pixel-level information. We simply consider bounding box-level annotations as a coarse mask for each object to supervise our method. The proposed module can be effortlessly integrated into any region-based detector to boost detection. Extensive experiments on ImageNet VID and EPIC KITCHENS datasets demonstrate consistent and significant improvement when we plug our BoxMask module into numerous recent state-of-the-art methods.

1. Introduction

With the recent advancements in deep convolutional neural networks [32, 61, 56], object detection in still images has gained a remarkable progress [23, 47, 44, 52, 21]. The naive idea of applying image-based detectors on each frame to perform Video Object Detection (VOD) often underperforms, owing to the deteriorated object appearance due to motion blur, rare poses, and part occlusions in videos. Therefore, exploiting the encoded temporal information in videos [67, 68, 58, 24] has become a de facto choice to tackle these challenges.

Earlier video object detection techniques utilizing temporal information mainly operate under two paradigms. The first category of methods applies post-processing on temporal information to make still image object detection results [30, 36, 35, 3] more consistent and stable. Alternatively, the second group leverages the feature aggregation of temporal information [67, 8, 58, 63, 11, 24]. Albeit these region-based state-of-the-art systems have greatly boosted the performance of VOD, they suffer from differentiating

the confusing objects with similar appearances or uniform motion attributes.

We observe that most of the previous approaches [67, 58, 24, 11] operate on instance-level feature aggregation that imminently neglects the refined pixel-level representation, resulting in acceptable localization but inferior classification. As illustrated in the first two rows of Figure 1, although the object detector exploits spatio-temporal context from support frames ($t - s$ and $t + s$) to refine proposal features, it produces false positives by classifying background as a *Bear* and misclassifies *Watercraft* with a *Car* at the target frame t . To overcome this hurdle, we design a novel module called BoxMask that exploits class-aware pixel-level temporal information to boost VOD. Inspired by [31] in still images, the BoxMask predicts a class-aware segmentation mask for each region of interest along with the conventional classification and localization. Since this paper deals with the problem of object detection in videos, we investigate bounding box-level annotation to generate coarse masks which supervise our BoxMask network. The advantages of adopting our BoxMask head are two folds. First, the class-aware pixel-level features reduce the hard false positives between objects with low spatial and temporal inter-class variance. Second, since the size of the predicted mask is identical to the target region, fine-grained pixel-level learning assists the detector in precise localization. We summarize the main contribution of this paper as follows:

- We observe that object misclassification is the crucial obstacle that limits the upper bound of existing video object detection methods. We further revisit the idea of leveraging bounding box annotations to supervise both regression and mask prediction (see Figure 1).
- We propose BoxMask, an extremely simple yet effective module that learns additional discriminative representations by incorporating class-aware pixel-level information to boost VOD.
- Our BoxMask is a plug-and-play module and can be integrated into any region-based detection method.

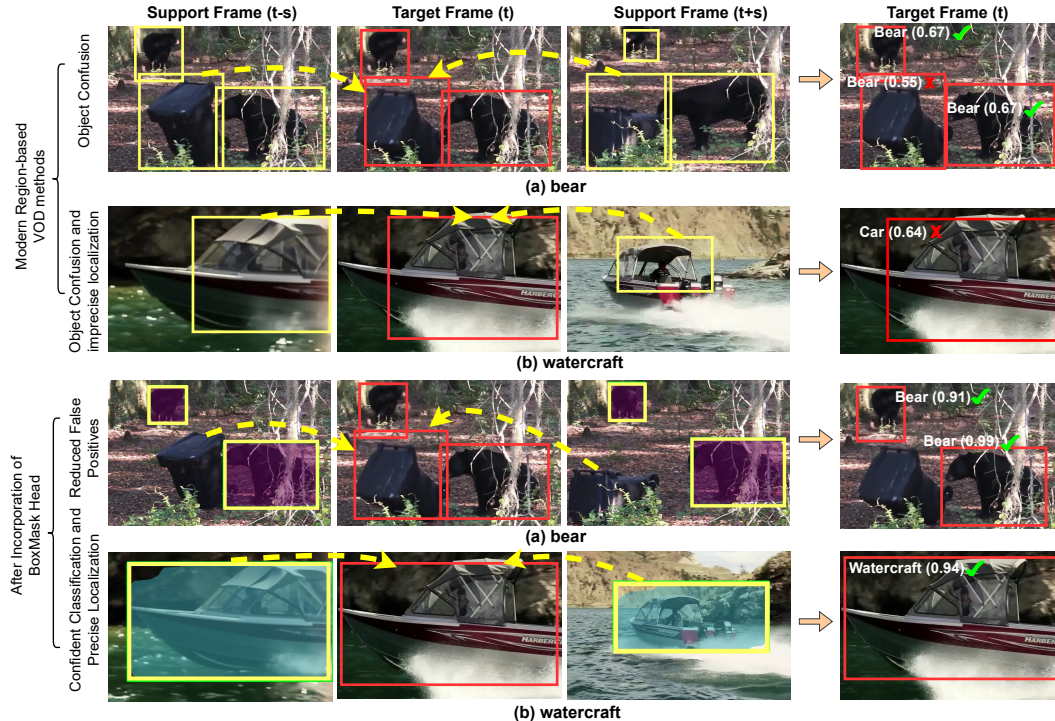


Figure 1. Motivation. Despite leveraging spatio-temporal information from support frames $t - s$ and $t + s$, modern VOD methods often misclassify objects with similar appearance and uniform motion characteristics. For instance, a moving object in the background is categorized as a *bear* in (a), while *Watercraft* is mistaken for a *Car* in (b). To address this, we devise a simple BoxMask module that learns pixel-level features by introducing crucial discriminative cues to boost detection among confused object categories. Note that with fine-grained pixel-level learning, our BoxMask removes misclassification of background in (a) and correctly categorizes *Watercraft* in (b). Best view it on the screen.

With our novel class-aware pixel-level learning introduced in recent state-of-the-art methods, we achieve an absolute gain of 1.8% in mAP and 2.1% in mAP on ImageNet VID and EPIC KITCHENS benchmarks, respectively.

2. Related Work

Object Detection in Images. The existing methods in image-based object detection can be mainly divided into single-stage detectors [42, 44, 45, 46, 9, 21] and multi-stage or region-based detectors [47, 6, 7, 26, 34]. Mask R-CNN [31] replaces RoI Pooling with RoIAlign and introduces an extra instance segmentation head that not only improves instance segmentation but advances object detection. Cheng et al. [12] blame the weak classification head for inferior detections and propose to ensemble the classification scores of Faster R-CNN [47] and R-CNN [23] as a remedy. IoU-Net [33] proposes a separate confidence mechanism for localization. Double-Head R-CNN [59] disentangles the detection head by treating classification with the fully connected head and regression with a convolution head. Along with this direction, seminal work [51] incorporates TSD in a region-based detector [47] that learns different features for classification and regression. Later, separate losses are

added to the whole loss function to optimize detection. Similar to these works in still images [31, 59, 51, 33], we observe that a naive sibling head in the region-based detector [47] confuses objects with similar motion characteristics and leads to sub-optimal video object detection.

Box-supervised Semantic and Instance Segmentation in Images. There has been an increasing trend in exploiting bounding box annotations to enhance weakly supervised instance and semantic segmentation approaches in still images [14, 39, 37, 41, 4]. The main reason is that bounding boxes contain knowledge about the precise location of each object, and they are approximately 35 times faster to annotate than per-pixel labeling [19, 2]. Along with a similar direction, our work exploits box-level annotations to generate coarse masks, eventually boosting video object detection.

Object Detection in Videos. Prior methods for video object detection have two directions. One direction exploits the redundancy in video frames by incorporating optical flow [68, 65], scale-time lattice [8], reinforcement learning capabilities [63], and heatmaps [62] to reduce the cost of the feature extraction process by propagating keyframe features to other frames in videos. Another line of work leverages temporal information encoded in videos to boost VOD, and our work operates on this trend. Existing techniques ex-

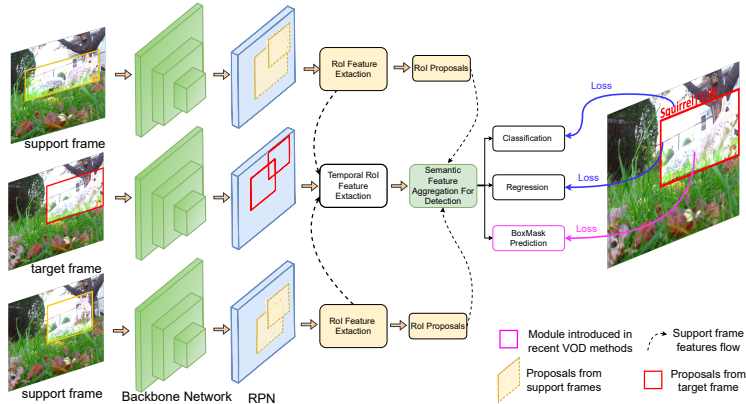


Figure 2. Architectural overview of modern region-based VOD methods and our proposed modules highlighted in magenta. Alongside spatio-temporal features, our method introduces important class-aware pixel-level features, which effectively tackles object confusion to boost performance in modern region-based video object detection methods.

exploit temporal information in two ways. The first way is to refine the detection results with post-processing methods [30, 36, 35]. Although these approaches improve the performance of VOD, they heavily rely on the image-based detector trained with no knowledge of temporal information. On the contrary, The second direction is to capitalize temporal information during the training stage [67, 57, 20, 5, 25, 50, 58, 17, 68, 65, 8, 60, 16, 53, 10, 11, 24, 64]. Some of these methods utilize optical flow [18] to warp and aggregate features across frames [67, 57, 35]. Despite the improvement, the optical flow based-methods fail in the case of occlusions. Most existing region-based VOD methods [67, 67, 58, 24] tackle the inherent challenges by aggregating temporal features. However, they mainly rely on instance-level feature aggregation, which pays less attention to the content of object proposals, resulting in confusion between objects with similar appearance and motion characteristics. Very recently, TransVOD [64] introduces the transformer-based VOD method by extending the Deformable DETR [66] with a temporal transformer to aggregate object queries from different video frames.

Tackling Object Confusion in Videos. Han et al. [29] are the first to highlight object confusion as to the main problem in VOD. They propose exploiting inter-video and intra-video proposal relations to tackle object confusion. Another seminal works [27, 28] attempts to solve this problem by devising better feature aggregation schemes that enhance target frame feature representation. Despite the gratifying improvement in detection, these approaches rely on a region-based detector that focuses more on discriminating between background and foreground regions than differentiating between various foreground regions [12]. Moreover, these methods operate on complex pipelines to produce impressive results. Alternatively, we design a simple but effective BoxMask module that achieves similar performance upon integrating into recent region-based VOD methods.

3. Method

This section first describes an overview of the modern region-based detectors in VOD by diving into the inherent misclassification problem in Section 3.1. Later, we explain the proposed BoxMask module and its learning mechanism in Sections 3.2 and 3.3, respectively.

3.1. Revisiting Region-based Detectors in VOD

Figure 2 depicts an overview of region-based detectors in VOD. First, a backbone network extracts spatial features from the target frame (the actual frame on which detection needs to be executed) and support frames (other video frames that assist the detection on a target frame). Subsequently, a Region Proposal Network (RPN) [47] predicts object proposals for each frame and aims to minimize the regression loss L_{reg} and classification loss L_{cls} defined as:

$$L_{rpn} = L_{cls}(p, p^*) + p^* \cdot L_{reg}(t, t^*) \quad (1)$$

where p is the estimated probability of a proposal being an object and p^* represents 1 or 0 depending upon the label of the anchor box. The term t denotes the coordinates of the predicted object proposal, and t^* is the ground truth. Here, note that the classification loss L_{cls} in Equation 1 only focuses on improving the objectness of proposals instead of object classification.

In the second stage, feature aggregation is performed between object proposal features of the target frame and support frames in a video. These aggregated features are pooled by an RoI Align pooling operator and propagated to the detection head designed to optimize multi-class classification and regression. For training, the detection loss is given by:

$$L_{det} = L_{cls}(p_c, y) + L_{reg}(t, t^*) \quad (2)$$

where p_c represents the predicted class distribution and y is the class label of an object in a target frame. For comprehensive details about the parameterization of RPN and

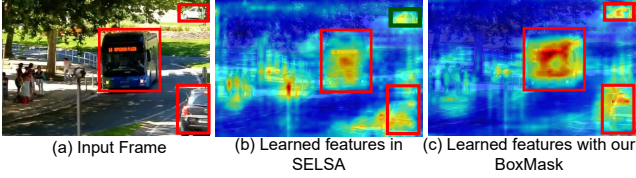


Figure 3. Visualization of learned class activation maps of SELSA and SELSA+BoxMask. (a) shows a sample input frame with target bounding boxes in red. (b) highlights that existing instance-level aggregation methods like SELSA [58] inevitably pay less attention to blurred and partly occluded objects, leading to missed detections (highlighted in green). (c) depicts that our fine-grained pixel-level learning brings additional discriminative cues that enrich target object features and suppress needless features.

detection head, we refer readers to [47]. Since the optimization of these region-based detectors relies on the cumulative sum ($L_{cls} + L_{reg}$), it converges to a compromising sub-optimal of two tasks [12]. Consequently, despite aggregating object proposal features from several support frames, the performance of current state-of-the-art VOD methods degrades due to the underlying object confusion caused by similar appearance and uniform motion characteristics. Furthermore, most existing methods operate on instance-level feature aggregation, which ignores blurred and partly occluded instances, leading to missed detection, as illustrated in Figure 3(b). In this paper, we aim to alleviate these limitations by incorporating class-aware pixel-level information in the detection head that brings additional discriminative features. To this end, we propose BoxMask, which assists the optimization of the detection head by enhancing target object features and discouraging irrelevant features, as visualized in Figure 3(c).

3.2. BoxMask

After extracting object proposals by RPN, in the second stage, we also have a set of aggregated proposal features $O = \{o_k\}_{k=1}^K$ from target and support frames, where K is the number of proposals. The BoxMask head predicts a binary mask for each RoI for classification and regression. Note that contrary to predicting instance mask in Mask R-CNN [31], our method predicts the mask of a complete bounding box along with classification and localization to simplify the overall multi-stage pipeline. Figure 4 visualizes the integration of the BoxMask module in region-based VOD methods.

Temporal RoI Feature Extraction for BoxMask Head. RoIAlign [31] pooling has outsmarted the RoIPool [22] operation to extract feature maps for each RoI, and it has been widely used in recent state-of-the-art VOD methods [58, 29]. Instead of the conventional RoIAlign operation, our method follows the spirit of [24] and exploits temporal information to extract RoI features. Given a group of video frames $\{V_{t+s}\}_{s=-N/2}^{N/2}$ and corresponding feature

maps $\{F_{t+s}\}_{s=-N/2}^{N/2}$ generated from the backbone network, where F_t represents feature maps of the target frame, and F_{t+s} ($s \neq 0$) denotes feature maps of support frames. First, we extract RoI features of target frame R_t by applying conventional RoIAlign on the target frame proposals, and target frame feature maps F_t . Then, in order to extract the most similar support frame RoI features R_{t+s} for target frame RoI features R_t , we compute the cosine similarity $C_{t+s} \in \mathbb{R}^{H \times W}$ between support frame feature maps F_{t+s} and a target frame RoI features R_t as follows:

$$C_{t+s} = R_t \otimes \{F_{t+s}\}^T \quad (3)$$

where \otimes represents matrix multiplication, and $\{\cdot\}^T$ highlights the matrix transposition. Later, analogous to [24], we employ multi-head self-attention [55] to aggregate target RoI features R_t and support frame RoI features R_{t+s} to form temporal RoI features for the target frame \bar{R}_t :

$$\bar{R}_t = MSA(R_t, R_{t+s}) \quad (4)$$

where MSA is a multi-head self-attention operation [55]. An overview of temporal RoI feature extraction is depicted in Figure 2. We refer readers to [24] for the detailed parameterization of temporal attentional feature aggregation of RoI features.

Instance Feature Extraction and Prediction. The BoxMask head is a fully convolutional [43] instance segmentation head in which, first, the temporal RoIAlign operation extracts 14×14 RoI features that are propagated into a single 3×3 convolutional layer to learn instance features. Contrarily to the complex instance segmentation problem [31, 7], we aim to predict the pixel mask of a rectangular bounding box. Therefore, we empirically establish that a single convolutional network is an optimal choice (see Section 4.4). As depicted in Figure 4, our prediction head contains a 2×2 deconvolution with a stride of 2, followed by the 1×1 convolution that predicts an output mask of size $C \cdot (m \times m)$ for each RoI, where C represents a total number of classes and $(m \times m)$ is the resolution.

3.3. Learning and Optimization

To alleviate the problem of object confusion and imprecision localization in videos, we view detection as a pixel-level classification problem. Furthermore, since our method operates in an end-to-end manner, it is robust to various datasets and backbone networks.

Generating Ground Truth. Considering our work deals with video object detection, an accurate object mask annotation is not available. Therefore, we revisit the exploitation of bounding boxes [14, 41, 39] in VOD and generate a mask with the given bounding box annotations to supervise the BoxMask head. Given the ground truth of bounding boxes represented by $B_{box} \in \mathbb{R}^{K \times 5}$, where K denotes

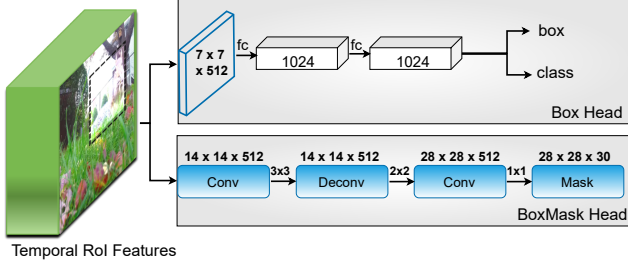


Figure 4. This figure presents the overall architecture of the detection phase equipped with a BoxMask head at the bottom. Numbers on blocks represent spatial resolution and channels, whereas numbers on arrows are the size of the kernel. In the BoxMask head, Conv and Deconv denote convolutions and deconvolutions, respectively. Thanks to its simplistic design, the proposed BoxMask head can be integrated into any region-based VOD method.

the set of bounding boxes consisting of 4 coordinates along with a corresponding class label. We define the bounding box mask tensor as $M_{box} \in \mathbb{R}^{m \times (L+1)}$, where m is the predicted spatial resolution and $(L+1)$ denotes L object classes and the background. We create a bounding box mask tensor M_{box} by labeling all the pixels inside the box with a corresponding class label. Following [37, 39], if two boxes overlap, we consider that a smaller box is in the front and label the pixels with a class of the smaller bounding box. The remaining pixels, not packed in any bounding box, belong to the background class.

BoxMask Loss and Multi-Task Learning. We employ bounding box mask tensor M_{box} to optimize the mask prediction by minimizing the cross-entropy loss L_{bm} :

$$L_{bm} = -\frac{1}{m} \sum_{c=0}^L \sum_{i=1}^m M(i, c) \log(y(i, c)) \quad (5)$$

where L_{bm} allows the network to predict the class for each pixel in each sampled RoI. This decouples the prediction of mask and class labels. Moreover, it assists the feature learning for localization because the predicted mask aims to be proportionally identical to the target bounding box. Upon integrating the BoxMask head in region-based video object detection methods, the detection loss explained in Equation 2 becomes

$$L_{det} = L_{cls} + L_{reg} + \lambda L_{bm} \quad (6)$$

where λ is the hyperparameter to control the weight of the BoxMask loss. We empirically set $\lambda=0.5$ in all experiments unless stated otherwise (refer to Section 2 in supplementary materials).

4. Experiments and Results

4.1. Experimental Setup

Datasets and Evaluation Metrics. We perform extensive experiments on the ImageNet VID dataset [49]. The dataset

comprises 3862 training videos and 555 validation videos with labeled bounding boxes of 30 classes. Following existing methods [58, 24, 67, 68], we train our models on the intersection of ImageNet DET and VID datasets[49] by utilizing the split provided in [67]. For direct comparison with prior works, we validate our models on the ImageNet VID validation set by using mean average precision (mAP) as a metric.

Training and Inference Details. We employ ResNet-50 [32] as the backbone network for ablation studies. In addition to ResNet-50, we utilize more powerful ResNet-101 [32] and ResNeXt-101 [61] to compare performance with existing methods. The backbone networks are initialized with ImageNet [38] pre-trained weights. We use SGD to train our models on 7 epochs with a total batch size of 8 on 8 GPUs. The training starts with an initial learning rate of 0.01, which is divided by 10 at the 4-th and 6-th epoch. For direct comparison, we sample one training frame (target frame) and two random frames (support frames) from the same video. During inference, we sample T frames (support frames) from the same video in addition to the target frame. Adopting [5, 24], we replicate the first/last frame of the video if support frames exceed the video start/end. Since our method detects objects in a target frame, the BoxMask module is switched off during the inference. Analogous to prior works [24, 58, 67], Non-Maximum Suppression (NMS) with an IoU threshold of 0.5 is incorporated to reduce reduplicate detections. The frames are resized to a shorter side of 600 pixels during both training and inference. For a detailed summarization of network architecture, refer to supplementary material (Section 1).

4.2. Effect of BoxMask on ImageNet VID Benchmarks

We compare performance between state-of-the-art systems equipped with our BoxMask module and summarize the results in Table 1. For a fair comparison, we reproduce the results of recent methods [11, 13, 24, 58, 67] by utilizing the original code from the authors. Therefore, for TF-Blender [13], we include results with their module crafted in FGFA [67]. Looking at the results in Table 1, our proposed BoxMask brings consistent and significant gains when incorporated into existing state-of-the-art methods with all three backbones. When BoxMask is plugged into TROI [24], we accomplish new state-of-the-art results with 80.7% mAP on the ResNet-50 backbone. Furthermore, leveraging our BoxMask module, all methods [11, 13, 24, 58, 67] with similar backbones enjoy gains from 0.4% (ResNeXt-101) to 1.8% (ResNet-50) in mAP. We argue that prior feature aggregation methods heavily rely on the capabilities of backbone networks, which results in inferior performance on a relatively weaker backbone of ResNet-50. Alternatively, our pixel-level feature informa-

Methods	mAP(%)	mAP(%)	mAP(%)
	R-50	R-101	RX-101
SFB _{NIPS'15} [47]	70.1	74.1	76.4
FGFA _{ICCV'17} [67]	74.7	77.8	79.6
SELSA _{ICCV'19} [58]	78.4	80.2	83.1
MEGA _{CVPR'20} [11]	77.3	81.6	-
TF-Blender _{ICCV'21} [13]	75.4	79.3	80.1
TROI _{AAAI'21} [24]	78.9	82.0	84.3
SFB + BoxMask	71.2 _{↑1.1}	75.0 _{↑0.9}	77.2 _{↑0.8}
FGFA + BoxMask	75.6 _{↑0.9}	78.7 _{↑0.7}	80.0 _{↑0.4}
SELSA + BoxMask	79.5_{↑1.1}	81.1 _{↑0.9}	83.5 _{↑0.4}
MEGA + BoxMask	78.2 _{↑0.9}	82.3_{↑0.7}	-
TF-Blender + BoxMask	76.3 _{↑0.9}	79.9 _{↑0.6}	80.4 _{↑0.3}
TROI + BoxMask	80.7_{↑1.8}	83.2_{↑1.2}	84.8_{↑0.5}

Table 1. Comparison with existing state-of-the-art methods on the ImageNet VID dataset. The SFB represents Single-Frame Baseline, Faster R-CNN, utilized as a base detector in all experiments. R and RX denote ResNet and ResNeXt backbone networks. The two best results are highlighted in red and blue.

tion in the BoxMask complements existing temporal feature aggregation schemes in [58, 24], obtaining superior gains in performance.

4.3. Qualitative Analysis

Visual Detection Results. Figure 5 illustrates the detection results of two recent state-of-the-art methods integrated with our BoxMask module in odd and even rows, respectively. We can see that SELSA [58] yields false negatives (turtle as a background) and false positives (turtle as a bird) in the case of rare poses in (a). On the other hand, these false detections are reduced with the introduction of our BoxMask module. Similarly, in the case of motion blur and part occlusion, our method alleviates misclassification (watercraft as a car by TROI [24]) by learning fine-grained pixel-level temporal information. These results show that adopting the BoxMask module in region-based VOD methods introduces class-aware pixel-level feature aggregation across different video frames that facilitates VOD under challenging conditions. Refer to supplementary material for more qualitative analysis.

Visual Proposal Feature Analysis. Following [29], we extract the learned proposal features before classification on the target frame and visualize them with t-SNE in Figure 6. We can see that proposal features of SELSA misclassifies proposals into incorrect clusters. For instance, proposals of watercraft and a bus incorrectly fall into the cluster of a car due to similar appearance and motion characteristics. Alternatively, when BoxMask is integrated into SELSA [58], we observe that proposal features of confusing object categories are clearly separated from each other. The main reason is that pixel-level feature aggregation enables the network to correctly distinguish proposals by decreasing the

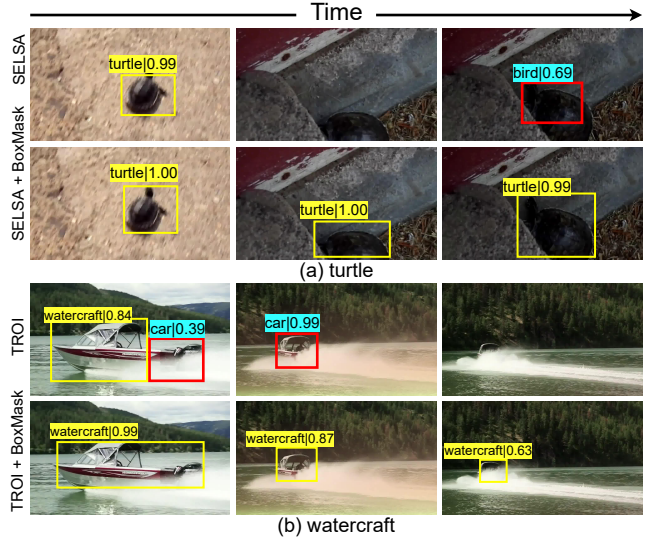


Figure 5. Qualitative analysis of existing methods with and without the BoxMask module integrated into the ImageNet VID dataset under different scenarios. Clearly, our BoxMask module facilitates SELSA [58] and TROI [24] to alleviate misclassification and imprecise localization in case of rare pose ((a) turtle), motion blur, and part occlusion ((b) watercraft), respectively. Best view it on the screen and zoom in.

intra-class and increasing the inter-class variance.

Analysis on Object Categories. Since our work mainly alleviates the object confusion in videos, we compare the performance in terms of mAP per category between the modern RoI-based VOD method [58] incorporated with and without our BoxMask module. We present the top 5 most improved classes and the top 5 most worsened categories in Figure 7. It is evident that the introduction of our pixel-level feature learning produces significant performance gains in motorcycle, domestic_cat, and cattle. The reason is that these objects have low inter-class variance due to similar appearance and motion characteristics. The pixel-level learning in our BoxMask effectively tackles this challenge and improves overall performance, as illustrated in Figure 5.

4.4. Ablation Studies

Sampling Support Frames. We follow the spirits of [58] and [24] to analyse the influence of number of frames and sampling strides during testing. Moreover, we examine the number of support frames sampled over an entire video. Figure 8(a) exhibits the influence of an increasing number of support frames T . We start with a single-frame detector by setting the frame stride S to 1. The mAP improves with the increasing number of frames, and it tends to stabilize at 74.4 mAP at $T = 26$. Later, we set T to 26 and start increasing the frame stride S . As illustrated in Figure 8(b), the mAP consistently improves with rising stride and eventually settles at $S = 7$. Finally, to exploit the whole video information, we make S adaptive to the length of the

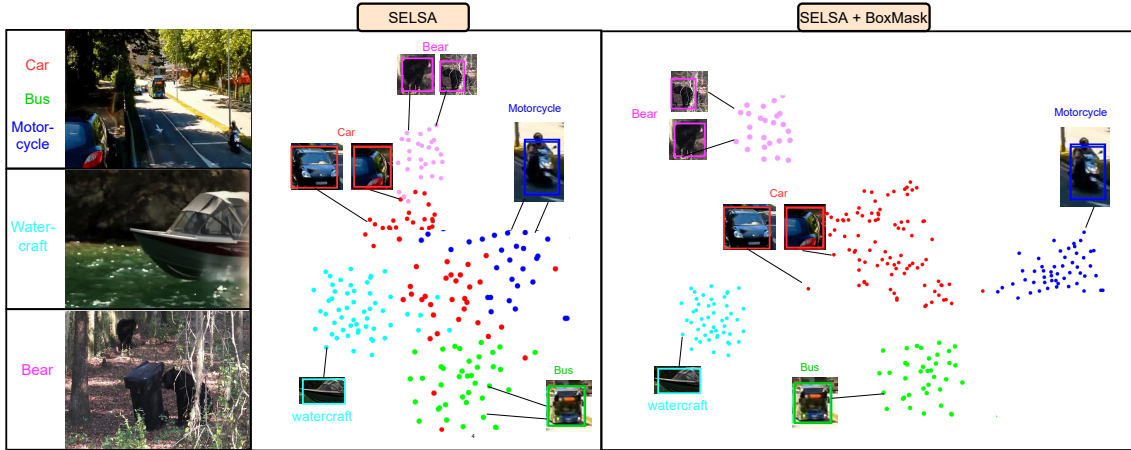


Figure 6. t-SNE visualization of learned proposal features with and without our BoxMask module in SELSA [58]. With instance-level feature aggregation only in SELSA, proposals of objects with similar motion characteristics (*Bus*, *car*, and *Watercraft*) mistakenly fall into each other’s cluster. Our class-aware pixel-level learning in BoxMask introduces discriminative cues which alleviate this object confusion, as shown in SELSA+BoxMask. Best view in color. For the complete figure with all 30 categories, refer to supplementary material.

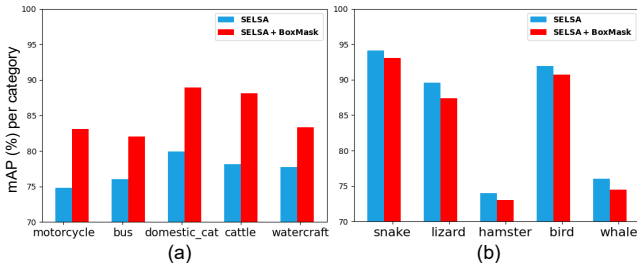


Figure 7. Performance comparison in terms of mAP per category. Subplots (a) and (b) denote the five most improved and most dropped classes when the BoxMask is equipped in SELSA [58].

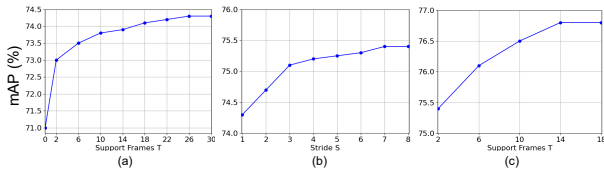


Figure 8. Ablation studies for frame sampling methods. (a) Investigating the effect of the different number of frames by fixing frame stride to 1. (b) Examining the effect of different frame strides by fixing the number of support frames to 26. (c) Assessing the effect of a different number of support frames sampled over the complete video.

video corresponding to the number of support frames T . Figure 8(c) shows that by leveraging only 2 support frames sampled over the entire video, mAP of 75.4 is achieved, surpassing the mAP on 26 succeeding support frames. The performance further boosts with the rise in the number of support frames and finally stabilizes at 14. We use the uniform sampling method with $T = 14$ in all the experiments unless stated otherwise.

Effectiveness of BoxMask. To investigate the flexibility and effectiveness of our method, we reproduce 4 existing ROI-based VOD methods [24, 58, 67, 68] and incorporate

Methods	AP _{0.5} (%)	AP _{0.75} (%)	AP _{0.5:0.95} (%)	Runtime(FPS)
DFE [68]	70.3	45.7	42.7	58.8
FGFA [67]	74.7	52.0	47.1	19.3
SELSA [58]	78.4	52.5	48.6	13.9
TROI [24]	78.9	52.8	48.6	7.3
DFE + BoxMask	71.3 _{↑1.0}	47.7 _{↑2.0}	44.6 _{↑1.9}	51.8 _{↓8.0}
FGFA + BoxMask	75.6 _{↑0.9}	54.2 _{↑2.2}	49.4 _{↑2.3}	17.3 _{↓2.0}
SELSA + BoxMask	79.5 _{↑1.1}	55.7 _{↑3.2}	50.0 _{↑2.0}	13.0 _{↓0.9}
TROI + BoxMask	80.7 _{↑1.8}	57.8 _{↑5.0}	51.8 _{↑3.2}	7.2 _{↓0.1}

Table 2. Tradeoff between effectiveness and efficiency of the proposed BoxMask using ResNet-50 as the backbone network. The run time is tested on a single DGX A100 GPU.

our BoxMask module. Table 2 summarizes the trade-off between the effectiveness and efficiency of the proposed BoxMask module. Looking at Table 2, we observe regular and substantial improvements in all 4 methods when equipped with our simple yet effective BoxMask module. Adopting BoxMask in TROI, we achieve a mAP@0.5 of 80.7(%), the new state-of-the-art result on ResNet-50. Similarly, on an increasing IoU threshold of 0.75, we notice a significant gain of 5 points in mAP when BoxMask is equipped with TROI. This reflects that our pixel-level learning not only alleviates object confusion but also yields high-quality predictions.

Effect on increasing convolutions. We investigate the design of our BoxMask network prior to pixel-wise class prediction. Specifically, for the instance feature extraction head, we study the impact of an increasing number of 3×3 convolutional layers, N_c . As shown in Table 3, the mAP declines with the rise in N_c . We argue that there are two main reasons for such behaviour. First, since our BoxMask aims to predict a mask of a rectangular target object, an increasing number of convolutional layers introduce unnecessary complex parameters that lead to overfitting. Second, given

that our BoxMask is supervised on bounding box annotations (containing object and background), increasing the size of N_c allows the network to learn needless high-level features, which causes object confusion.

Size of RoI Features. As explained in Section 3.2, we perform Temporal RoiAlign to extract RoI features with spatial resolution of 7×7 . The RoI features are then upsampled to size 14×14 . Furthermore, we investigate different resolution settings for RoI and upsampled features for completeness. As summarized in Table 4, we accomplish an optimal trade-off between performance and efficiency upon setting the RoI and upsampled features size as 7 and 14, respectively.

(N_c)	mAP(%)	FPS	Params	RoI Upsample	mAP(%)	FPS
1	80.7	7.2	0.8	7	7	78.3
2	80.5	6.8	1.5	7	14	80.7
3	80.4	6.6	2.1	7	28	79.4
4	80.2	6.2	2.6	14	14	80.7

Table 3. Effect on increasing number of convolutional layers (N_c) in BoxMask. Params represents number of parameters $\times 10^6$.

Table 4. Effect on increasing size of RoI Features.

Computational Analysis. For brevity, we present the real-time performance of our BoxMask module in Table 2. We can observe that the speed of SELSA and SELSA+BoxMask are 13.9 and 13.0 FPS on a single DGX A100 GPU, respectively. Moreover, when our method is adopted in TROI [24], the speed drops by 0.1 FPS while achieving an mAP gain of 1.8 points. This demonstrates that the BoxMask module brings significant performance gains with a negligible increase in computation.

4.5. Additional Experiments on EPIC KITCHENS

Dataset and Implementation Details. Along with ImageNet VID dataset, we evaluate our method on a far more challenging EPIC KITCHENS dataset [15]. The VOD task in this dataset comprises 32 unique kitchens, including 290 classes. We employ 272 video sequences captured in 28 kitchens for training, whereas, for evaluation, 106 sequences are collected in the same 28 Kitchens (S1), and 54 sequences are gathered from other 4 unseen kitchens (S2). For direct comparison with prior works [24, 58], we adopt identical implementation settings explained in [58].

Performance Analysis. We reimplement prior works [58, 24] on the EPIC KITCHENS dataset and evaluate the results on an IoU threshold of 0.5 and 0.75. As summarized in Table 5, we observe consistent and significant performance gains when the proposed BoxMask is equipped in SELSA [58] and TROI [24]. It is important to emphasize

Methods	AP _{0.50} (S1)	AP _{0.75} (S1)	AP _{0.50} (S2)	AP _{0.75} (S2)
SELSA [58]	38.8	10.2	36.7	9.2
TROI [24]	42.2	13.3	39.6	11.3
SELSA + BoxMask	40.7 \uparrow _{1.9}	14.7 \uparrow _{4.5}	38.1 \uparrow _{1.4}	12.8 \uparrow _{3.6}
TROI + BoxMask	44.3 \uparrow _{2.1}	18.5 \uparrow _{5.2}	41.3 \uparrow _{1.7}	15.7 \uparrow _{4.4}

Table 5. Performance comparison without and with the BoxMask module in previous state-of-the-art methods on EPIC KITCHENS test set. S1 and S2 represent Seen and Unseen splits, respectively.

that on a higher IoU threshold of 0.75, our BoxMask further improves the mAP to (4.5/3.6) points for SELSA and (5.2/4.4) points for TROI for Seen/Unseen splits. This establishes that incorporating our simple BoxMask module in region-based detectors can boost the performance of VOD even on complex datasets.

4.6. Limitations

Albeit integrating our proposed module in VOD systems substantially improves detections, we notice that the performance drops for some object categories, as illustrated in Figure 7(b). We observe that our method yields false negatives (confusing objects with background) and false positives (confusing background with an object class). Such behaviour is due to the supervision from faulty object masks with no information on object boundaries. Therefore, our BoxMask network treats the background part in the bounding box as an object mask by learning class-aware pixel-level information. Thus, applying methods like GrabCut [48] and MCG [1] on bounding boxes to reduce the background content in object masks is one possible way to tackle this problem. Moreover, learning the refined instance mask from the coarse BoxMask in a weakly supervised manner as done in still images [40, 54] will introduce refined object boundaries, alleviating confusion between foreground and background pixels.

5. Conclusion

In this paper, we address the crucial problem of object confusion that limits the upper bound of video object detection models and present a simple yet effective BoxMask module as a remedy. Our method introduces class-aware pixel-level information that brings crucial discriminative indicators that enhance classification and localization. The proposed module is conceptually simple and can be applied to any region-based detection method to boost performance. Extensive experiments on ImageNet VID and EPIC KITCHENS datasets demonstrate that the introduction of our proposed method brings consistent and significant performance gains in recent video object detection methods.

References

- [1] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014.
- [2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What’s the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016.
- [3] Hatem Belhassen, Heng Zhang, Virginie Fresse, and El-Bay Bourennane. Improving video object detection by seq-bbox matching. In *VISIGRAPP (5: VISAPP)*, pages 226–233, 2019.
- [4] Míriam Bellver Bueno, Amaia Salvador Aguilera, Jordi Torres Viñals, and Xavier Giró Nieto. Budget-aware semi-supervised semantic and instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019*, pages 93–102, 2019.
- [5] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 331–346, 2018.
- [6] 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.
- [7] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiao-xiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019.
- [8] 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.
- [9] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, and Jian Sun. You only look one-level feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13039–13048, 2021.
- [10] Xingyu Chen, Junzhi Yu, Shihan Kong, Zhengxing Wu, and Li Wen. Joint anchor-feature refinement for real-time accurate object detection in images and videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(2):594–607, 2020.
- [11] 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.
- [12] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. In *Proceedings of the European conference on computer vision (ECCV)*, pages 453–468, 2018.
- [13] 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.
- [14] Jifeng Dai, Kaiming He, and Jian Sun. Boxesup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015.
- [15] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.
- [16] 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.
- [17] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7023–7032, 2019.
- [18] 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.
- [19] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [20] 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.
- [21] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.
- [22] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [23] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [24] 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.
- [25] 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.
- [26] Chaoux Guo, Bin Fan, Qian Zhang, Shiming Xiang, and Chunhong Pan. Augfpn: Improving multi-scale feature learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12595–12604, 2020.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yunying Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 784–799, 2018.
- [34] Chenhan Jiang, Hang Xu, Wei Zhang, Xiaodan Liang, and Zhenguo Li. Sp-nas: Serial-to-parallel backbone search for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11863–11872, 2020.
- [35] 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.
- [36] 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.
- [37] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017.
- [38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [39] Viveka Kulharia, Siddhartha Chandra, Amit Agrawal, Philip Torr, and Amrith Tyagi. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *European Conference on Computer Vision*, pages 290–308. Springer, 2020.
- [40] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3406–3416, 2021.
- [41] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652, 2021.
- [42] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [43] 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.
- [44] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [45] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.
- [46] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [47] 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.
- [48] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. ” grabcut” interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.
- [49] 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.
- [50] 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.
- [51] Guanglu Song, Yu Liu, and Xiaogang Wang. Revisiting the sibling head in object detector. In *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11563–11572, 2020.
- [52] 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.
- [53] Peng Tang, Chunyu Wang, Xinggang Wang, Wenyu Liu, Wenjun Zeng, and Jingdong Wang. Object detection in videos by high quality object linking. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1272–1278, 2019.
- [54] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021.
- [55] 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.
- [56] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.
- [57] 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.
- [58] 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.
- [59] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10186–10195, 2020.
- [60] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 485–501, 2018.
- [61] 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.
- [62] Zhujun Xu, Emir Hrustic, and Damien Vivet. Centernet heatmap propagation for real-time video object detection. In *European conference on computer vision*, pages 220–234. Springer, 2020.
- [63] Chun-Han Yao, Chen Fang, Xiaohui Shen, Yangyue Wan, and Ming-Hsuan Yang. Video object detection via object-level temporal aggregation. In *European conference on computer vision*, pages 160–177. Springer, 2020.
- [64] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvot: End-to-end video object detection with spatial-temporal transformers. *arXiv preprint arXiv:2201.05047*, 2022.
- [65] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7210–7218, 2018.
- [66] 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.
- [67] 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.
- [68] 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.