BoxMask: Revisiting Bounding Box Supervision for Video Object Detection

Khurram Azeem Hashmi  Alain Pagani  Didier Stricker  Muhammad Zeshan Afzal
DFKI - German Research Center for Artificial Intelligence, Kaiserslautern
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. The code will be available at https://github.com/khurramHashmi/BoxMask.

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 an 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 \((t - s \text{ and } t + s)\) 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.
Support Frame (t-s)

Support Frame (t+s)

Target Frame (t)

Object Confusion

imprecise localization

Car (0.64)

Object Confusion

Bear (0.94)

Bear (0.67)

(a) bear

Bear (0.55)

(b) watercraft

Bear (0.99)

(b) watercraft

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).

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-
Tackling Object Confusion in VOD. 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 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.

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.

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 \( L_{\text{reg}} \) and classification loss \( L_{\text{cls}} \) defined as:

\[
L_{\text{rpn}} = L_{\text{cls}}(p, p^*) + p^* L_{\text{reg}}(t, t^*)
\]

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 \( L_{\text{cls}} \) 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:

\[
L_{\text{det}} = L_{\text{cls}}(p_c, y) + L_{\text{reg}}(t, t^*)
\]

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
R. Learned features in
ation, our method follows the spirit of [24] and exploits
been widely used in recent state-of-the-art VOD meth-
eration to extract feature maps for each RoI, and it has
RoIAlign [31] pooling has outsmarted the RoIPool [22] op-
VOD methods.

Figure 4 visual-
bounding box along with classification and localization to
CNN [31], our method predicts the mask of a complete
Note that contrary to predicting instance mask in Mask R-
is the number of proposals. The BoxMask head predicts a
mask with the given bounding box annotations to super-
Figure 3. Visualization of learned class activation maps of SELSA
(a) Input Frame (b) Learned features in
SELSA (c) Learned features with our
BoxMask

detection head, we refer readers to [47]. Since the optimiza-
tion of these region-based detectors relies on the cumulative
sum \((L_{cla} + L_{reg})\), it converges to a compromising subop-
timal of two tasks [12]. Consequently, despite aggregating
object proposal features from several support frames, the
performance of current state-of-the-art VOD methods de-
grades due to the underlying object confusion caused by
similar appearance and uniform motion characteristics. Fur-
thermore, most existing methods operate on instance-level
feature aggregation, which ignores blurred and partly oc-
cluded instances, leading to missed detection, as illustrated
in Figure 3(b). In this paper, we aim to alleviate these limi-
tations 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 vi-

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 visual-
izes 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] op-
operation to extract feature maps for each RoI, and it has
been widely used in recent state-of-the-art VOD meth-
ods [58, 29]. Instead of the conventional RoIAlign oper-
ation, 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 con-
ventional RoIAlign on the target frame proposals, and tar-
get 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:

\[
C_{t+s} = R_t \odot \{F_{t+s}\}^T
\]

where \(\odot\) represents matrix multiplication, and \(\{\}\^T\) high-
lights 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 \(\mathcal{T}_t\):

\[
\mathcal{T}_t = MSA(R_t, R_{t+s})
\]

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 paramete-
ization of temporal attentional feature aggregation of RoI
features.

Instance Feature Extraction and Prediction. The Box-
Mask head is a fully convolutional [43] instance segmen-
tation head in which, first, the temporal RoIAlign oper-
ation extracts \(14 \times 14\) RoI features that are propagated
into a single \(3 \times 3\) convolutional layer to learn instance
features. Contrarily to the complex instance segmentation
problem [31, 7], we aim to predict the pixel mask of a rect-
angular 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 \(2 \times 2\) deconvolution with a stride of 2, followed
by the \(1 \times 1\) convolution that predicts an output mask of size
\(C (m \times m)\) 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 impre-
cision 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 anno-
tation is not available. Therefore, we revisit the exploita-
tion of bounding boxes [14, 41, 39] in VOD and generate
a mask with the given bounding box annotations to sup-
ervise the BoxMask head. Given the ground truth of bound-
ing boxes represented by \(B_{box} \in \mathbb{R}^{K \times 5}\), where \(K\) denotes
the set of bounding boxes consisting of 4 coordinates along
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.

with a corresponding class label. We define the bounding box mask tensor as \( M_{\text{box}} \in \mathbb{R}^{m \times (L+1)} \), where \( m \) is the predicted spatial resolution and \( L+1 \) denotes \( L \) object classes and the background. We create a bounding box mask tensor \( M_{\text{box}} \) 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 \( M_{\text{box}} \) to optimize the mask prediction by minimizing the cross-entropy loss \( L_{bm} \):

\[
L_{bm} = -\frac{1}{m} \sum_{c=0}^{L} \sum_{i=1}^{m} \log(y(i, c))
\]

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

\[
L_{det} = L_{cls} + L_{reg} + \lambda L_{bm}
\]

where \( \lambda \) 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 ResNetXt-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-
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 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 video corresponding to the number of support frames $T$.
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.

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 our BoxMask module. Table 2 summarizes the trade-off between the 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.

![Figure 8(c)](https://example.com/image.png)

Table 2. Tradeoff between 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 $3 \times 3$ 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 annota-
tions (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 RoI Align to extract RoI features with spatial resolution of $7 \times 7$. The RoI features are then upsampled to size $14 \times 14$. 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.

<table>
<thead>
<tr>
<th>$(N_c)$</th>
<th>mAP(%)</th>
<th>FPS</th>
<th>Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80.7</td>
<td>7.2</td>
<td>0.8</td>
</tr>
<tr>
<td>2</td>
<td>80.5</td>
<td>6.8</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>80.4</td>
<td>6.6</td>
<td>2.1</td>
</tr>
<tr>
<td>4</td>
<td>80.2</td>
<td>6.2</td>
<td>2.6</td>
</tr>
</tbody>
</table>

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

**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 that on a higher IoU threshold of 0.75, our BoxMask fur-

<table>
<thead>
<tr>
<th>Methods</th>
<th>$AP_{0.50}(S1)$</th>
<th>$AP_{0.75}(S1)$</th>
<th>$AP_{0.50}(S2)$</th>
<th>$AP_{0.75}(S2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELSA [58]</td>
<td>38.8</td>
<td>10.2</td>
<td>36.7</td>
<td>9.2</td>
</tr>
<tr>
<td>TROI [24]</td>
<td>42.2</td>
<td>13.3</td>
<td>39.6</td>
<td>11.3</td>
</tr>
<tr>
<td>SELSA + BoxMask</td>
<td>40.7</td>
<td>14.7</td>
<td>38.1</td>
<td>12.8</td>
</tr>
<tr>
<td>TROI + BoxMask</td>
<td>44.3</td>
<td>18.5</td>
<td>41.3</td>
<td>15.7</td>
</tr>
</tbody>
</table>

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.

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.

**Acknowledgement.** This work was partially funded by the Federal Ministry of Education and Research Germany under the project DECODE (011W21001) and partially by the European project INFINITY (883293).
References


