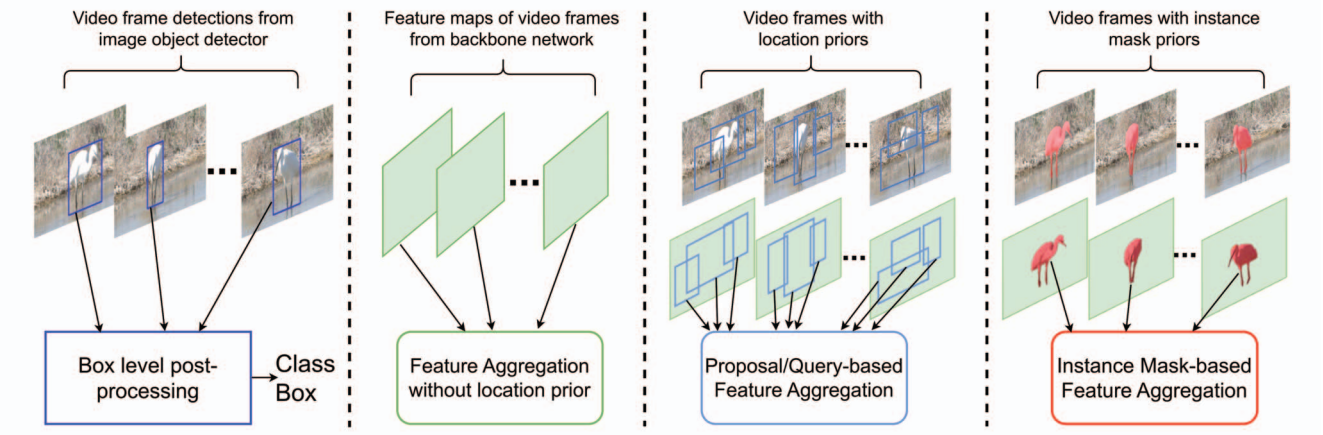
Beyond Boxes: Mask-Guided Spatio-Temporal Feature Aggregation for Video Object Detection

Khurram Azeem Hashmi* Talha Uddin Sheikh* Didier Stricker Muhammad Zeshan Afzal
DFKI - German Research Center for Artificial Intelligence, Kaiserslautern
firstname[0]_firstname[1].lastname@dfki.de



(a) Box-level: Seq-NMS [25] **(b) Frame-level**: FGFA [75] **(c) Proposal-level**: SELSA [65] **(d) Instance Mask-level**: Ours

Figure 1. Evolution of exploiting temporal information in video object detection. (a) Box-level post-processing to refine detections. **(b)** Feature aggregation across entire video frames. **(c)** Temporal feature aggregation guided by region-location priors from each frame. **(d)** Our instance mask-based aggregation refines the focus to instance boundaries, reducing background noise and improving feature aggregation.

Abstract

The primary challenge in Video Object Detection (VOD) is effectively exploiting temporal information to enhance object representations. Traditional strategies, such as aggregating region proposals, often suffer from feature variance due to the inclusion of background information. We introduce a novel ***instance mask-based feature aggregation*** *approach, significantly refining this process and deepening the understanding of object dynamics across video frames. We present* ***FAIM****, a new VOD method that enhances temporal* ***F****eature* ***A****ggregation by leveraging* ***I****nstance* ***M****ask features. In particular, we propose the lightweight Instance Feature Extraction Module (IFEM) to learn instance mask features and the Temporal Instance Classification Aggregation Module (TICAM) to aggregate instance mask and classification features across video frames. Using YOLOX as a base detector, FAIM achieves 87.9% mAP on the ImageNet VID dataset at 33 FPS on a single 2080Ti GPU, setting a new benchmark for the speed-accuracy trade-off. Additional experiments on multiple datasets validate that our approach is robust, method-agnostic, and effective in multi-object tracking, demonstrating its broader applicability to video understanding tasks.*

1. Introduction

Video Object Detection (VOD) aims to identify and locate objects in a video sequence. It has numerous applications, including autonomous driving and video surveillance [41, 42, 60]. Despite remarkable progress in object detection [4, 11, 13, 21, 29, 33, 34, 48–50, 57, 57, 69, 74], applying image-based detectors [11, 20, 48, 50, 74] to individual video frames often results in decreased performance. This decline is due to degradation from motion blur, rare poses, camera defocus, and occlusions [75]. However, video frames have the advantage of temporal context, as the detection in one frame

[1]* Equal technical contribution.

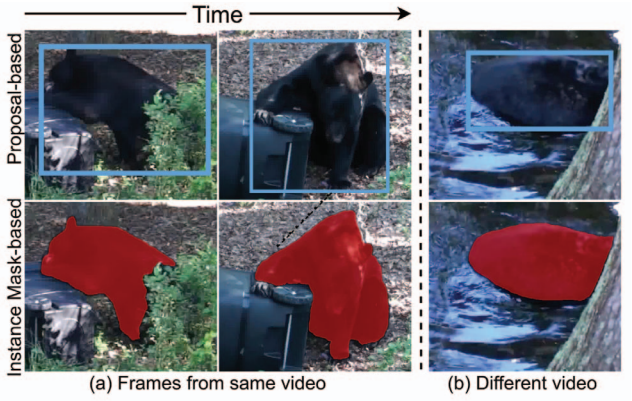


Figure 2. Exploiting temporal information in proposal-based feature aggregation in **blue** against our instance mask-based feature aggregation method in **red** for the class **Bear**. Leveraging instance mask-level information significantly reduces variance among **Bear** proposals within and across videos.

can leverage information from surrounding frames. Thus, *effectively exploiting the temporal information in videos* is crucial for addressing the challenges of VOD.

The exploration of temporal information in VOD has evolved significantly, as illustrated in Fig. 1. Starting with box-level post-processing [3, 19, 25, 36, 37], progressing through image-level feature aggregation [10,56,61,64,73,75, 76], and culminating in proposal (query)-level feature aggregation [5,9,16,22,24,26,27,31,35,43,46,54,55,63,65,67,71]. This progression highlights a critical evolution toward enhanced computational efficiency and detection accuracy, focusing on proposal or query-level feature aggregation to reduce background noise and intra-class feature variance compared to image-level aggregation [65]. However, this approach remains sub-optimal, as it includes background features, amplifying intra-class variance, particularly in occlusion scenarios (see Fig. 2). This limitation remains a fundamental bottleneck in VOD.

Building on this shift and its limitations, we pose the question: *Can we improve VOD by refining proposal/box-level information to the instance mask-level during temporal feature aggregation*? This paper introduces a novel paradigm in video object detection: instance mask-based feature aggregation. Unlike established proposal-level feature aggregation (Fig. 1c), our approach shown in Fig. 1d, leverages instance pixel-level features to aggregate temporal features across video frames. By focusing on the most granular level—directly around object instances—this method effectively minimizes background noise and intra-class feature variance, as depicted in Fig. 2. Inspired by recent advancements in box-based semi-supervised instance segmentation [6, 39, 58], we introduce the lightweight Instance Feature Extraction Module (IFEM) to learn instance mask features. The Feature Prediction Selection Module (FPSM) refines these features and forwards them to our Temporal Instance and Classification Aggregation Module (TICAM) for final predictions. Additionally, the instance mask features from IFEM are optimized using a mask loss function, comparing them with the pseudo ground truth mask obtained from any box-based instance segmentation methods like Box2Mask [39] or SAM [38].

Based on these modules, we present **FAIM**, a new end-to-end VOD framework that enhances temporal **F**eature **A**ggregation through leveraging **I**nstance **M**ask features. Following YOLOV [54], FAIM extends YOLOX [20] to include the learning of video object and instance mask features with minimal modifications, as shown in Fig. 4. Our instance mask-based feature aggregation through FAIM achieves the best speed and accuracy trade-off, as shown in Fig. 3. To summarize, our main contributions are:

1) Paradigm shift: We introduce a novel paradigm of instance mask-based feature aggregation in VOD, significantly refining the aggregation process and offering a deeper understanding of object dynamics across video frames.

2) FAIM: The proposed modules in FAIM, such as **IFEM and TICAM**, are **method-independent** and can be adapted to other VOD approaches to improve performance (Table 6).

3) Robustness and Generalizability: Extensive experiments validate that our approach is robust (Tables 7 and 8) and applicable to different video understanding tasks, including **multi-object tracking** (see Table 9).

2. Related Work

Box-level Post-Processing. Early efforts in video object detection (VOD) [3, 19, 25, 36, 37] primarily utilized temporal information through box-level post-processing strategies (Fig. 1a). In these approaches, conventional image object detection methods [11, 21, 48, 50] are first applied to individual frames. The predictions from these frames are then refined by integrating temporal cues across sequences. This is achieved through various techniques, including tubelet proposals [37], tracking [19], Soft-NMS [3], and re-scoring of detections during Non-Maximum Suppression (NMS) [25]. While these methods have shown improvements, they do not leverage temporal context during the training phase. Consequently, inaccuracies in initial frame-level detections can propagate throughout the sequence, impacting the overall performance.

Frame-level Feature Aggregation. Frame-level feature aggregation represents a more sophisticated approach for leveraging temporal information in VOD [10, 61, 64, 73, 75, 76] (Fig. 1b). This methodology begins with feature extraction using backbone networks like ResNet [30] and Swin Transformer [44], followed by aggregating these features over a temporal window to boost their discriminative capability for the target frame. Pioneering works such as DFF [76] and FGFA [75] employ optical flow fields [18] to align and aggregate features from adjacent frames. Subsequent ad-

vancements have focused on improved feature propagation methods [64, 73] and the effective integration of temporal and spatial features [10, 61]. While effective, these methods often overlook long-term temporal dependencies and can be computationally demanding due to frame-level feature aggregation. To overcome these challenges, we propose instance mask-based feature aggregation that not only exploits pixel-level features but also limits the feature aggregation from the image to the instance level.

Proposal/Query-level Feature Aggregation. Recent advancements in video object detection (VOD) methods [5, 16, 22, 24, 26, 27, 31, 35, 55, 63, 65, 67, 71] have explored aggregating features at the object proposal level (Fig. 1c). This approach provides a more context-sensitive and computationally efficient method for incorporating temporal cues. For instance, SELSA [65], a pioneering work, aggregates proposal features among video frames based on semantic similarity. TROIA [22] and MEGA [5] utilize temporal information for extracting and enhancing proposal features, with MEGA introducing a memory-based mechanism to exploit both local and global information. MAMBA [55] introduces a pixel or instance-level memory bank to optimize memory updates for each frame. These methods typically generate proposals on each video frame using a region proposal network [50]. Additionally, query-based feature aggregation methods [9, 31, 63, 71], utilizing Transformer-based detectors like Deformable DETR [74], have been explored. Very recently, YOLOV [54] has emerged as a state-of-the-art approach in VOD, balancing speed and accuracy effectively. YOLOV treats detections from a powerful single-stage detector such as YOLOX [20] as proposals and aggregates features among video frames for final results. These developments highlight the significant gains brought by focusing on objects while aggregating temporal information in VOD. However, all these methods are limited to optimizing box-level information surrounding the object due to the absence of object masks. In contrast, this paper proposes a novel paradigm of instance mask-based feature aggregation, focusing on fine-grained object-level information.

3. Method

Overview. This section first outlines a straightforward approach to transform any proposal-based feature aggregation method to our instance mask-based framework in § 3.1. Building on these principles, we then delve into the design decisions behind FAIM, detailing its unique architectural elements and functionalities in § 3.2.

3.1. From Proposal to Instance Mask-Based Feature Aggregation

Proposal-based Feature Aggregation. Let us recall the proposal-based feature aggregation scheme in video object detection [27, 54, 65, 71]. Given an m frames $\{I_1, I_2, \ldots, I_m\}$

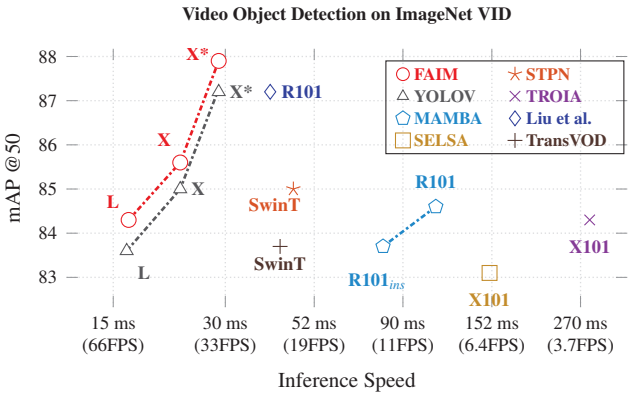


Figure 3. Speed and accuracy Trade-off. FAIM outperforms prior state-of-the-art methods on the ImageNet VID benchmark. Besides QueryProp, MAMBA, and Liu *et al.*, all results are reported on the 2080Ti GPU. * denotes results with post-processing.

from the same video, we first extract feature maps for each frame using a shared backbone network $b_{\text{cnn}}(\cdot;.\theta_{\text{cnn}})$. The feature map F_t for the frame I_t is obtained as: $F_t = b_{\text{cnn}}(I_t;\theta_{\text{cnn}})$, where θ_{cnn} are the parameters of the backbone network. For each frame I_t, a set of proposals $\{P_{t1}, P_{t2}, \ldots, P_{tn_t}\}$ is generated using a Region Proposal Network (RPN) [50] or some image detector [20, 74], where n_t is the number of proposals for frame I_t. RoIAlign [29] is then applied to extract proposal features X_{tj} for each proposal P_{tj} from the feature map F_t:

$$X_{tj} = \text{RoIAlign}(F_t, P_{tj}). \tag{1}$$

These proposal features $\{X_{t1}, X_{t2}, \ldots, X_{tn_t}\}$ for frame I_t are then aggregated with proposal features from other frames to enhance the feature representation as follows:

$$X_{\text{agg}} = \mathscr{A}(\{X_{1j}, X_{2j}, \ldots, X_{mj}\}_{j=1}^{n_t}), \tag{2}$$

where n_t is the number of proposals for the frame I_t. The aggregation function $\mathscr{A}(\cdot)$ can be a mean, max, or a more complex function like attention-based [59] feature aggregation. Before aggregation, these proposal features (used in Eq. 2) are calibrated across space-time based on semantic similarity [22, 65], object classes [23], memory [5, 55], or prediction confidence [54]. Despite progress, a major limitation is that each proposal feature X_{tj} contains background features in the bounding box. Appearance degradation (common in videos due to rare poses or camera defocus) adversely affects feature aggregation, increasing the intra-class feature variance and decreasing the inter-class feature variance for objects with similar backgrounds. We illustrate this limitation in Fig 2. To overcome this, we propose using instance mask-based features instead of region proposals during the spatio-temporal feature aggregation. This simple modification isolates object features from the background, reducing intra-class feature variance.

Instance Mask-based Feature Aggregation. Let us consider the same example. After obtaining the RoI features X_{tj}

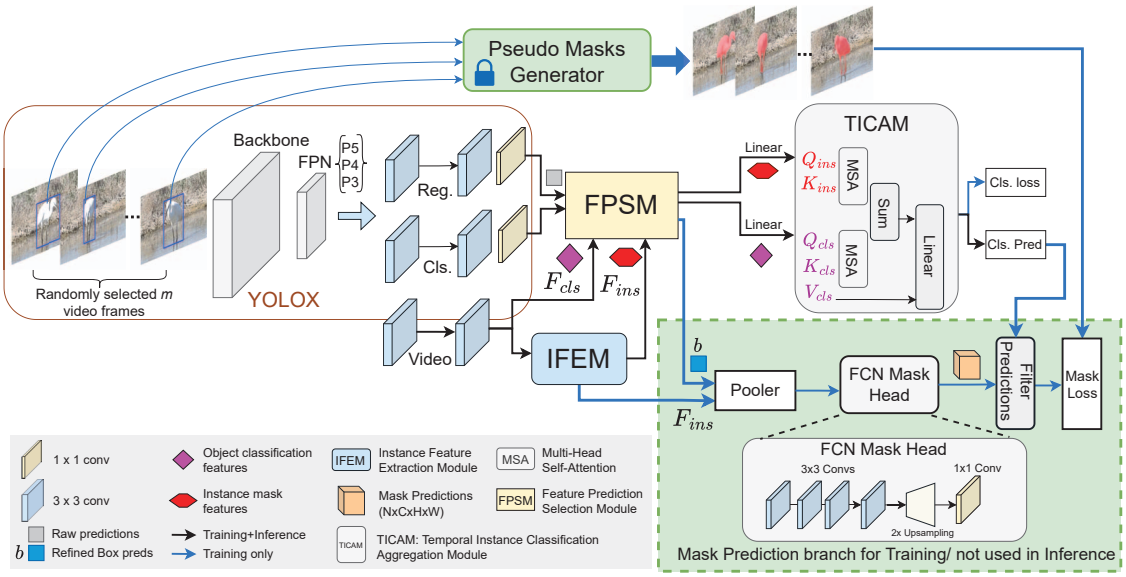


Figure 4. Overview of FAIM framework.Randomly sampled frames from a video are input into YOLOX [20] for initial feature extraction and prediction using multi-scale features (P3-P5). The IFEM processes video object features to produce instance mask features (Eq. 3), while the FPSM filters the features for object classification. IFEM's instance mask features and FPSM's refined predictions are combined to predict instance masks, which are optimized against pseudo-ground truth masks. The learned instance mask features and classification features are then fed into the TICAM for final classification. **Inference:** Components in green are excluded during inference. However, IFEM continues to provide high-quality instance mask features, enhancing feature aggregation in TICAM for robust predictions.

for each proposal P_{tj} using Eq. 1, we propose an Instance Feature Extraction Module (IFEM) to distill instance mask features M_{tj} from proposal features X_{tj}:

$$M_{tj} = \text{IFEM}(X_{tj}) \tag{3}$$

These instance-mask features M_{tj} are used to predict instance masks (using fully-convolutional head [11, 29]). Then, these predicted masks are compared against pseudo ground truth masks generated by any box-based instance segmentation methods like Box2Mask [39] or a zero-shot segmentation model like SAM [38]. This comparison refines the instance mask features M_{tj}, optimizing them to align closely with the pseudo ground truth masks, thus enhancing the quality of the instance-specific representation. These instance-mask features can then replace the proposal features in Eq. 2 as:

$$M_{\text{agg}} = \mathscr{A}(\{M_{1j}, M_{2j}, \ldots, M_{mj}\}_{j=1}^{n_t}) \tag{4}$$

This ensures feature aggregation with a higher level of granularity, focusing on the object instances and reducing the background noise. Thus, *thanks to this simple recipe, any proposal-based feature aggregation scheme can be converted to the instance mask-based feature aggregation approach, without hand-annotated mask labels.* Following this recipe, we propose FAIM to verify the effectiveness of our proposed instance mask-based feature aggregation in VOD.

3.2. FAIM

The FAIM (illustrated in Fig. 4) incorporates the instance mask-based feature aggregation in Video Object Detection (VOD). Motivated by the impressive real-time performance of YOLOV [54], FAIM employs YOLOX [20] as a base detector with minimal modifications to achieve impressive performance while making it attractive for real-time applications. We now detail each modification.

FPSM: Feature and Prediction Selection Module. Initial predictions from the YOLOX [20] detection head serve as region proposals $\{P_{t1}, P_{t2}, \ldots, P_{tn_t}\}$. Extracting and aggregating features from all these proposals increases computations. Therefore, FPSM filters proposal features and predictions effectively. Following common conventions [50, 54, 65], we select top k (e.g., $k = 750$) predictions based on the confidence scores and perform Non-Maximum Suppression (NMS) to obtain refined n ($n << k$) proposals. Next, to obtain video object-level features, we extend the neck of the base detector with the video object branch as depicted in Fig. 4. Similar to the classification and regression branch in [20], this branch contains two 3×3 convolutional layers. Unlike YOLOV [54], which employs the detector's regression features for feature aggregation, our video object branch decouples video object features into classification features F_{cls} and instance mask features F_{ins}. The F_{cls} are directly filtered based on the refined predictions in FPSM, whereas F_{ins} are first extracted by our proposed instance feature extraction module.

IFEM: Instance Feature Extraction Module. As depicted

in Fig. 4, the IFEM is a simple and lightweight module that projects the video object feature into the video instance mask feature space using a single 3×3 convolutional layer. Let $V^R \in \mathbb{R}^{H \times W \times C}$ represent the video object features extracted from our video object branch, where H, W, and C denote the height, width, and number of channels, respectively. Using Eq. 3, IFEM applies a convolution operation $C(\cdot;\theta_{\text{conv}})$ with parameters θ_{conv} to transform V^R into instance mask features $F_{ins} \in \mathbb{R}^{H \times W \times C'}$, with C' as the number of channels in the transformed feature space. During training, these instance mask features F_{ins} are utilized to predict instance masks, ensuring that the features highly represent the instance masks. Note that the implementation details of IFEM are not important, and even more advanced networks such as [7,40] can be employed. Later, similar to F_{cls}, we filter F_{ins} according to refined predictions in FPSM and feed them to the temporal instance classification aggregation module.

TICAM: Temporal Instance Classification Aggregation Module. Now that we have the filtered video instance mask features and video object classification features, we employ multi-head attention [59] to aggregate them as explained in Eq. 4. In our TICAM, the input to multi-head attention includes Q_{cls} and Q_{ins}, formed by stacking the features from the classification features F_{cls} and the instance mask features F_{ins} for all proposals across the temporal space (i.e., $Q_{cls} = \text{LP}([F_{cls1}, F_{cls2}, \ldots, F_{clsm}]^T)$ and $Q_{ins} = \text{LP}([F_{ins1}, F_{ins2}, \ldots, F_{insm}]^T)$. Here, $\text{LP}(\cdot)$ is the linear projection operator. To verify the effectiveness of the instance mask-based feature aggregation, we adopt the feature aggregation of YOLOV [54] to establish the direct comparison between YOLOV and our FAIM. However, in our TICAM, the temporal aggregation of object classification and instance mask features reduces the background information, producing more discriminative features for VOD. Moreover, it is important to emphasize that our TICAM is independent of the employed feature aggregation scheme. Thanks to Eqs. 3 and 4, it can incorporate other aggregation approaches [5, 10, 55].

Learning Instance Masks. Learning instance masks is a crucial step in our FAIM during training, as shown in Fig. 4. In the mask prediction branch, we pool the region features from F_{ins} (from IFEM), according to refined box predictions b (from FPSM), and feed them to Fully Convolutional Network (FCN) [11] to predict instance masks. Here, the Pooler is RoIAlign [29] as explained in Eq. 1. The FCN mask head contains four 3×3 convolutional layers, followed by upsampling and 1×1 convolution to predict mask $M \in \mathbb{R}^{N \times C \times H \times W}$ with N and C represent the number of proposals and total classes, respectively. H and W denote the size of the predicted mask. For each proposal, we generate C class-specific predictions. However, comparing $N \times C$ masks with G ground truth masks (where $G \ll N$) can be sub-optimal during loss computation. Therefore, we use the TICAM's classification outputs to select masks from M corresponding to positively classified proposals. Formally, let $P = \{p_1, p_2, \ldots, p_N\}$ be the set of proposals, and the classification predictions from TICAM for these proposals are denoted as $T = \{t_1, t_2, \ldots, t_N\}$, where $t_i \in \{1, 2, \ldots, C\}$ represents the predicted class for proposal p_i. The refined mask predictions M' are obtained by:

$$M' = \{m'_i \mid m'_i = M[i, t_i, :, :], \forall i \in \{1, 2, \ldots, N\}, \quad (5)$$

where, $m'_i \in \mathbb{R}^{H \times W}$ is the mask prediction for proposal p_i corresponding to its classified category t_i. The proposed filtration approach reduces the number of masks processed and focuses learning on class-specific features, enhancing the network's ability to distinguish between classes. We optimize refined mask predictions M' by minimizing the cross entropy loss, jointly trained in a multi-task fashion [29], along with detection losses from the base detector [20]. Again, it is worth mentioning that the implementation details of the mask prediction branch are not important. Here, the goal is not to predict the most accurate segmentation masks but to push F_{ins} to learn instance-specific features. Refer to Appendix A.2 for the mask loss computation.

Method	Source	Backbone	mAP(%)↑	Time (ms)↓
SELSA [65]	ICCV2019	X101	83.1	153.8
RDN [16]	ICCV2019	R101	81.8	162.6
MEGA [5]	CVPR2020	R101	82.9	230.4
TROIA [22]	AAAI2021	X101	84.3	285.7
MAMBA [55]	AAAI2021	R101	84.6	110.3*(T)*
QueryProp [28]	AAAI2022	R101	82.3	30.8*(T)*
SparseVOD [27]	BMVC2022	R101	81.9	142.4
FAQ [9]	CVPR2023	R50	81.7	163.2
Liu *et al*. [43]	ICCV2023	R101	87.2	39.6*(T)*
STPN [56]	ICCV2023	SwinT	85.0	45.7
TransVODLite [71]	TPAMI2022	SwinT	83.7	42.1
YOLOV-S [54]	AAAI2023	MCSP	77.3	11.3
YOLOV-L [54]		MCSP	83.6	16.3
YOLOV-X [54]		MCSP	85.0	22.7
FAIM-S	Ours	MCSP	**78.2**$_{+0.9}$	**11.6**
FAIM-L		MCSP	**84.3**$_{+0.7}$	**16.5**
FAIM-X		MCSP	**85.6**$_{+0.6}$	**22.7**
With Post-processing				
YOLOV-S [54]	AAAI2023	MCSP	80.1	11.3 + 6.9
YOLOV-L [54]		MCSP	86.2	16.3 + 6.9
YOLOV-X [54]		MCSP	87.2	22.7 + 6.1
FAIM-S	Ours	MCSP	**80.6**$_{+0.5}$	**11.6 + 6.9**
FAIM-L		MCSP	**87.0**$_{+0.8}$	**16.5 + 6.9**
FAIM-X		MCSP	**87.9**$_{+0.7}$	**22.7 + 6.9**

Table 1. Comparing accuracy and speed on the ImageNet VID dataset. T denotes the inference time from corresponding papers tested on a different GPU. MCSP is the Modified CSP v5 backbone adopted in YOLOX. Improvements in red highlight gains over YOLOV. **Our FAIM consistently outperforms YOLOV with all variants of YOLOX while maintaining comparable runtime**.

4. Experiments

Dataset and Evaluation Metrics. Our primary experiments are conducted on the ImageNet VID dataset [52], comprising 3,862 training videos and 555 validation videos, spanning 30 object classes with annotated bounding boxes. Adhering to standard VOD protocols [5,54,65,71], we utilize a combined

dataset of ImageNet VID and DET [52] for training and report results on the validation set using the mean average precision (mAP) metric. Inference runtime is reported in milliseconds (ms) on a single NVIDIA 2080Ti GPU unless stated otherwise.

Base Detector and Backbones. Consistent with prior works [20, 32, 71], we initialize our base detector using COCO pre-trained weights from YOLOX [20]. Our FAIM is evaluated across different YOLOX variants (YOLOX-S, YOLOX-L, YOLOX-X), each incorporating the Modified CSP v5 backbone [62]. Consequently, we refer to our FAIM variants as FAIM-S, FAIM-L, and FAIM-X.

Training. To directly compare with YOLOV [54], we adopt an identical training strategy and follow the original codebase[1] from the authors. We sample one-tenth of the frames from the ImageNet VID training set to address the redundancy. The base detectors are trained as in [54] with a batch size of 16 on 2 GPUs. When base detectors are integrated into our FAIM, we fine-tune them on a batch size of 16 on a single GPU. The same learning schedule is adopted, and only the newly added video object feature branch, instance feature extraction module, FCN mask head, and multi-head attention are fine-tuned. To generate pseudo ground truth instance masks, we try pre-trained SAM with the ViT-H [17] image encoder and pre-trained Box2Mask [39] with ResNet-101 [30] backbone network. Owing to better performance, we select ground truth instance masks from SAM for experiments. Refer to Appendix C.1 for the performance comparison between SAM and Box2Mask. During training, in the FPSM, the NMS is set to 0.75 to select box predictions and features. In TICAM, the number of frames m is set to 16.

Testing. During testing, the NMS threshold is set to 0.5, whereas the number of frames m for feature aggregation is empirically set to 32. Complete implementation details are provided in Appendix A.1.

4.1. Main Results

Our FAIM aims for real-time video object detection (VOD). Therefore, we mainly compare it with several state-of-the-art methods focussing on real-time VOD. As shown in Table 1, we present a quantitative analysis comparing both mAP (%) and the inference run-time of FAIM against other prominent VOD methods [5,9,16,22,27,28,43,54–56,65,71]. YOLOV, our direct competitor with the same detector and backbone, is compared in all three variants. Thanks to our novel instance mask-based feature aggregation and the efficiency of the single-stage detector, FAIM consistently and significantly surpasses the previous state-of-the-art, specifically YOLOV [54], achieving the highest mAP of 87.9% and 85.6% with and without sequential post-processing [53], respectively. Notably, our lightweight instance feature extraction module results in a negligible increase in inference run-time (+0.3 ms and +0.2 ms compared to YOLOV-S and YOLOV-L, respectively). However, it brings considerable gains of +0.9% and +0.7% in mAP without post-processing. When adopting a larger detector like YOLOX-X, the difference in run-time becomes negligible, while the mAP improvement remains significant at +0.7%. Apart from [28, 43, 55], all models are evaluated on the same GPU for a direct comparison. Moreover, it is worth noting that the proposed modules in FAIM are method-agnostic and can be plugged into other VOD methods to improve performance (see Table 6).

Qualitative Comparison. We extract and compare the proposal features from the FAIM's Temporal Instance Classification Aggregation Module (TICAM) and YOLOV's Feature Aggregation Module (FAM) [54] using t-SNE in Fig.5. As demonstrated, FAIM's use of instance mask-level features offers a significant advantage, as it leads to compact clustering of proposals within each class, reducing intra-class variance. Moreover, it increases the separation between different classes, particularly among visually similar or background-heavy categories, such as *Watercraft* and *Whale*. This improvement allows FAIM to better differentiate between objects with overlapping contexts or similar backgrounds. Appendix B offers more qualitative analysis.

4.2. Ablation Studies

We analyze the design decisions in FAIM using YOLOX-S as a base detector on the validation set of ImageNet VID [52] dataset. We employ similar settings to Sec. 4 and report performance on the standard mAP_{50} and runtime in milliseconds (ms). More ablations studies and evaluations are provided in Appendix C.

Effectiveness of each component. Table 2 analyzes the contributions of our proposed modules IFEM and TICAM on both YOLOV-S [54] and YOLOX-S [20]. For YOLOV-S, the baseline achieves a mAP of 77.3%. Adding IFEM results in a mAP increase of +0.6%, bringing the total to 77.9%, with a negligible runtime increase (+0.3ms). This improvement suggests that even without incorporating instance mask features into the feature aggregation module, the addition of instance mask learning in [54] helps refine the temporal object classification queries Q_{cls} for better classification. In YOLOX-S, originally a single-frame detector, adding both IFEM and TICAM transforms it into FAIM-S, a video object detection model. This yields a substantial improvement, increasing mAP from 69.5% to 78.2% (+8.9%) with a modest runtime increase of +2.2ms. IFEM introduces instance mask learning, while TICAM effectively aggregates temporal mask and classification features across frames, reducing feature variance and significantly improving detection performance. A comparison of TICAM and IFEM with standard attention-based methods [12] can be found in Appendix C.4.

Reference frame sampling. Consistent with previous re-

[1] https://github.com/YuHengsss/YOLOV

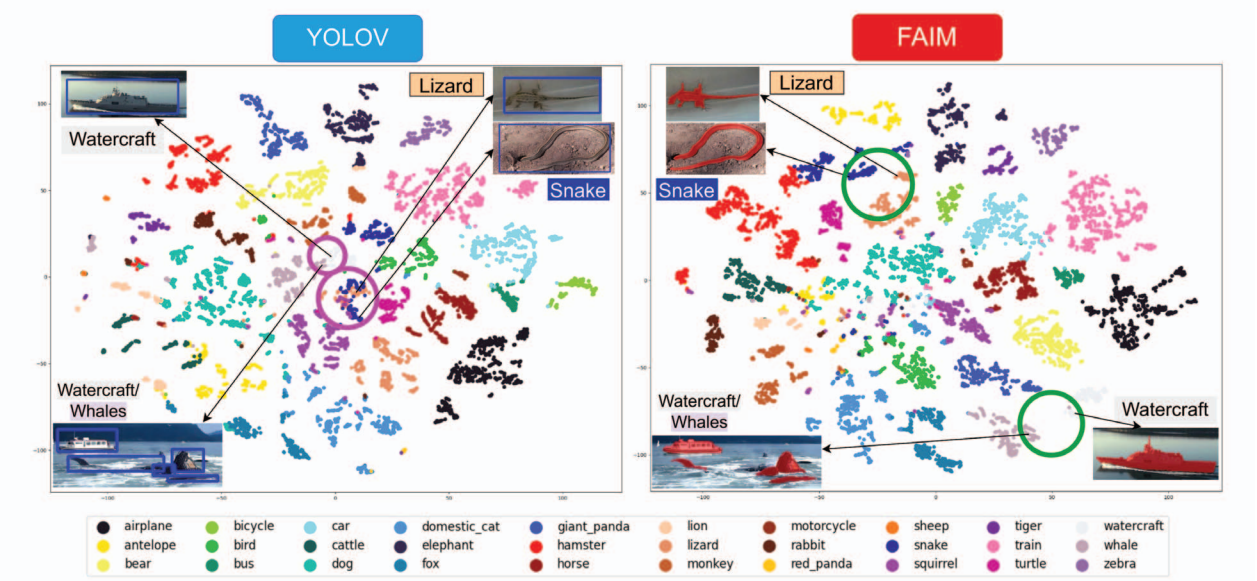


Figure 5. TSNE of proposal features from YOLOV [54] and FAIM on the ImageNet VID dataset. Feature confusion in YOLOV is marked with magenta circles ◯, and corrections in FAIM with green circles ◯. The blue bounding box □ shows the area used for feature aggregation in YOLOV, while FAIM uses the area in red mask. YOLOV confuses features between ***Snake*** and ***Lizard*** (highlighted with ◯), showing higher intra-class and lower inter-class variance due to background inclusion. FAIM's instance mask-based feature aggregation reduces this variance, forming clearer clusters. Similar improvements are seen with ***Watercraft*** and ***Whale***. Best viewed on a screen.

Method	IFEM	TICAM	T (ms)	mAP
YOLOV-S	✗	✗	11.30	77.3
YOLOV-S	✓	✗	11.60	77.9$_{+0.6}$
YOLOX-S	✗	✗	9.40	69.5
FAIM-S	✓	✓	11.60	**78.2**$_{+8.9}$

Table 2. Effectiveness of the modules proposed in FAIM.

$m_g \rightarrow$	3	7	15	23	31	39
mAP	75.4	76.8	77.7	77.9	**78.2**	78.2
$m_l \rightarrow$	3	7	15	23	31	39
mAP	71.8	72.6	73.4	73.8	74.3	74.6

Table 3. Varying global m_g and local reference frames m_l.

$n \rightarrow$	10	20	30	50	75	100
mAP	76.9	77.8	**78.2**	78.3	78.4	78.4
Time (ms)	10.68	10.98	**11.60**	14.17	20.02	30.08

Table 4. Investigating different number of proposals n in FPSM.

search [22,26,54,65], we explore both global and local frame sampling strategies in our work. The results in Table 3 reveal that using merely 3 global reference frames surpasses the performance achieved with 39 local reference frames. This finding is in line with prior works [22, 26, 54, 65]. Therefore, in alignment with the approach in [54], we adopt the global sampling strategy with m_l=31 as the default.

Number of Proposals. We study the effect of varying the number of prediction proposals n from 10 to 100 in FPSM. As shown in Table 4, our approach, FAIM, demonstrates a notable increase of 0.9% in mAP when n increases from 10 to 20. This performance already surpasses that of YOLOV-S [54] (with $n = 30$) by **+0.5%** in mAP, while also being faster by **0.6 milliseconds**. Further elevating n to 30 results in an additional mAP gain of +0.4%, albeit with an increase of 0.6 milliseconds in runtime. The improvement continues consistently as n is increased, reaching a plateau at $n = 75$. Given the quadratic complexity ($O(n^2)$) of the self-attention mechanism in TICAM, we opt for $n = 30$.

Design Choices for Mask Prediction. Table 5 presents an ablation study of the mask prediction branch in FAIM. We analyze the impact of pooling features from different scales (P3-P5) in the model's neck (see Fig. 4), as detailed in Eq. 1. Table 5a shows that pooling features from P5 yields the best mask prediction results. Hence, P5 is used by default. Table 5b explores varying the RoIAlign output size, with 32×32 chosen for optimal performance during training, as mask prediction is not required during inference. Table 5c demonstrates that filtering mask predictions based on TICAM's classification improves mAP by 0.4%. Table 5d shows that Binary Cross-Entropy (BCE) loss is the most effective for mask loss and is used by default. The mask prediction branch in FAIM is modular and can be fully modified. Further ablations are presented in Appendix C.

4.3. FAIM in other two-stage VOD Methods

Settings. Following the recipe detailed in § 3.1, this study evaluates the adaptability of our instance mask-based feature aggregation in two-stage, proposal-based VOD methodologies, namely SELSA [65] and TROIA [22]. Following the 1x schedule in MMTracking [8] with ResNet-50 as the backbone, we examine these methods with and without our instance mask-based feature aggregation scheme (see Eq. 3). Implementation details are available in Appendix A.2.

Results. Table 6 lists the results, demonstrating that the integration of instance mask-based feature aggregation yields a significant **improvement of more than 1% in mAP$_{50}$** for both SELSA [65] and TROIA [22]. Notably, these enhance-

Scale →	P3	P4	P5	P3-P5
mAP	78.1	77.9	**78.2**	78.2

(a) # FPN scale for RoIAlign.

Size →	14× 14	28 × 28	32×32
mAP	77.7	78.1	**78.2**

(b) RoIAlign output Size.

Mask loss→	Class Aware	Class Agnostic
mAP	**78.2**	77.8

(c) Instance Mask loss computation.

Loss →	Dice	BCE
mAP	77.9	**78.2**

(d) Loss Function.

Table 5. Ablating mask prediction branch in FAIM. Settings for results in § 4 are highlighted .

Method	mAP_{50}	mAP_{75}	$mAP_{50:95}$
SELSA* [65]	78.4	52.5	48.6
SELSA+Ours	**79.5**$_{+1.1}$	**54.4**$_{+1.9}$	**49.6**$_{+1.0}$
TROIA* [22]	78.9	52.8	48.8
TROIA+Ours	**80.1**$_{+1.2}$	**55.4**$_{+2.6}$	**50.0**$_{+1.2}$

Table 6. Exploring instance mask-based feature aggregation in other VOD methods. Results with * are reproduced. Improvement of over 1% is observed.

Method	AP50/AP75 (S1)	AP50/AP75 (S2)
Liu [43]	44.9/18.7	41.7/16.0
TROIA [22]	42.2/13.3	39.6/11.3
TROIA+Ours	**45.1/18.9** +2.9/+5.6	**42.0/16.2** +2.4/+4.9

Table 7. Results on EPIC KITCHENS-55 [14]. S1 and S2 are seen and unseen splits. We achieve new SOTA results.

Method	AP	AP50	AP75
YOLOV-X [54]	54.7	75.0	57.2
FAIM-X	**55.8** +1.1	**76.9** +1.9	**58.6** +1.4

Table 8. Our FAIM achieves **stronger gains of +1.9 points in AP50** on the OVIS [47] dataset with severe occlusions.

ments are achieved with minimal modifications, as detailed in § 4. These outcomes in Table 6 confirm the efficacy of the instance mask-based feature aggregation technique in two-stage proposal-based VOD methods, suggesting its potential for further improvements.

4.4. Additional VOD Benchmarks

Experiments on EPIC KITCHENS-55. Besides ImageNet VID, we report results on the more challenging EPIC KITCHENS-55 dataset [14], comprising ego-centric videos of 32 different kitchens and 290 classes. Implementation details are in Appendix A.3. Table 7 summarizes the results. When our proposed instance mask-based feature aggregation is integrated into TROIA [22], we surpass prior state-of-the-art results in both splits, affirming its applicability to challenging video object detection tasks.

Experiments on OVIS. Following the experimental setting in [54], we compare the performance of our FAIM and YOLOV on the Occluded Video Instance Segmentation (OVIS) dataset [47]. This dataset contains 25 classes and is notable for its high level of occlusion, with many objects being partially or fully occluded in multiple frames. Refer to Appendix A.4 for more implementation details. As shown in Table 8, FAIM-X surpasses YOLOV-X by a significant margin, highlighting the effectiveness and robustness of our instance mask-based feature aggregation on occluded VOD tasks.

Method	MOTA [2]↑	IDF1 [51]↑	HOTA [45]↑	IDS [2]↓
Tracktor* [1]	70.5	65.3	53.0	1442
Tracktor+Ours	**71.4**$_{+0.9}$	**66.7**$_{+1.4}$	**53.1**$_{+0.1}$	**1344**$_{-98}$
ByteTrack* [70]	86.4	82.7	65.5	995
ByteTrack+Ours	**88.1**$_{+1.7}$	**83.7**$_{+1.0}$	**68.9**$_{+3.4}$	**911**$_{-84}$

Table 9. Exploring instance mask-based learning in Multi-Object Tracking. Results with * are reproduced. Our method shows consistent gains across all metrics in both methods.

4.5. Application in Multi-Object Tracking (MOT)

Settings. Since consistent tracking and reidentification of objects is an important task in MOT, we experiment with two MOT methods (i.e. two-stage detector-based Traktor [1] and YOLOX-based ByteTrack [70]) and incorporate our instance-mask learning in the detector using Eq. 3. To validate the performance, we evaluate ByteTrack and Tracktor with and without our instance mask learning on the MOT20 [15] dataset. Complete details of experiments, dataset, and evaluation metrics are outlined in Appendix D.1.

Results. As summarized in Table 9, our proposed instance mask learning has significantly enhanced the performance of both Tracktor and ByteTrack across nearly all metrics. For instance, the MOTA score improves from **70.5 to 71.4** in Tracktor and from **86.4 to 88.1** in ByteTrack. These remarkable improvements suggest that exploiting instance mask information temporally not only enhances VOD but also significantly boosts MOT. Moreover, these findings outline the promising potential of our approach in other video understanding tasks [66, 68, 72]. Qualitative analysis is presented in Appendix D.2.

5. Conclusion and Discussion

This paper introduces a novel paradigm for video object detection through instance mask-based feature aggregation, refining the process to enhance object understanding across video frames. Extensive experiments on multiple benchmarks with different VOD and MOT methods validate our approach's effectiveness and highlight its potential to advance video understanding. Integrating instance mask learning into video understanding tasks opens novel research opportunities, especially when mask data is unavailable. Future work will explore unifying VOD, MOT, and video instance segmentation [66] into a cohesive framework.

Acknowledgements

This work was in parts supported by the EU Horizon Europe Framework under grant agreements 101135724 (LUMINOUS) and 101092312 (AIRISE).

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 8

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 8

[3] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms – improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2021. 1

[5] 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 (CVPR)*, June 2020. 2, 3, 5, 6

[6] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Qian Zhang, and Wenyu Liu. Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3145–3154, 2023. 2

[7] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, and Wenyu Liu. Sparse instance activation for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4433–4442, June 2022. 5

[8] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking, 2020. 7

[9] Yiming Cui. Feature aggregated queries for transformer-based video object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6365–6376, June 2023. 2, 3, 5, 6

[10] 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 (ICCV)*, pages 8138–8147, October 2021. 2, 3, 5

[11] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 1, 2, 4, 5

[12] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7373–7382, June 2021. 6

[13] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2988–2997, October 2021. 1

[14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 8

[15] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes, 2020. 8

[16] 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 (ICCV)*, October 2019. 2, 3, 5, 6

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6

[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. 2

[19] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[20] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1, 2, 3, 4, 5, 6

[21] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1, 2

[22] 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. 2, 3, 5, 6, 7, 8

[23] 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*, 32(12):8165–8178, 2022. 3

[24] 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. 2, 3

[25] 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. *CoRR*, abs/1602.08465, 2016. 1, 2

[26] Khurram Azeem Hashmi, Alain Pagani, Didier Stricker, and Muhammad Zeshan Afzal. Boxmask: Revisiting bounding box supervision for video object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2030–2040, January 2023. 2, 3, 7

[27] Khurram Azeem Hashmi, Didier Stricker, and Muhammamd Zeshan Afzal. Spatio-temporal learnable proposals for end-to-end video object detection. In *British Machine Vision Conference*, volume 2021, 2022. 2, 3, 5, 6

[28] Fei He, Naiyu Gao, Jian Jia, Xin Zhao, and Kaiqi Huang. Queryprop: Object query propagation for high-performance video object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):834–842, Jun. 2022. 5, 6

[29] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 3, 4, 5

[30] 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 (CVPR)*, June 2016. 2, 6

[31] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. End-to-end video object detection with spatial-temporal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 1507–1516, New York, NY, USA, 2021. Association for Computing Machinery. 2, 3

[32] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. End-to-end video object detection with spatial-temporal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1507–1516, 2021. 6

[33] Qinghang Hong, Fengming Liu, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dynamic sparse r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4723–4732, June 2022. 1

[34] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19702–19712, June 2023. 1

[35] Zhengkai Jiang, Yu Liu, Ceyuan Yang, Jihao Liu, Peng Gao, Qian Zhang, Shiming Xiang, and Chunhong Pan. Learning where to focus for efficient video object detection. In *European conference on computer vision*, pages 18–34. Springer, 2020. 2, 3

[36] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, and Wanli Ouyang. 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, 2018. 2

[37] 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 (CVPR)*, June 2016. 2

[38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2, 4

[39] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Risheng Yu Xiansheng Hua, and Lei Zhang. Box2mask: Box-supervised instance segmentation via level-set evolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 4, 6

[40] Zhanhao Liang and Yuhui Yuan. Mask frozen-detr: High quality instance segmentation with one gpu, 2023. 5

[41] Dongfang Liu, Yiming Cui, Yingjie Chen, Jiyong Zhang, and Bin Fan. Video object detection for autonomous driving: Motion-aid feature calibration. *Neurocomputing*, 409:1–11, 2020. 1

[42] Dongfang Liu, Yiming Cui, Xiaolei Guo, Wei Ding, Baijian Yang, and Yingjie Chen. Visual localization for autonomous driving: Mapping the accurate location in the city maze. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3170–3177. IEEE, 2021. 1

[43] Xin Liu, Fatemeh Karimi Nejadasl, Jan C. van Gemert, Olaf Booij, and Silvia L. Pintea. Objects do not disappear: Video object detection by single-frame object location anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6950–6961, October 2023. 2, 5, 6, 8

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 2

[45] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021. 8

[46] Shishir Muralidhara, Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. Attention-guided disentangled feature aggregation for video object detection. *Sensors*, 22(21):8583, 2022. 2

[47] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8):2022–2039, 2022. 8

[48] 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 (CVPR)*, June 2016. 1, 2

[49] Joseph Redmon and Ali Farhadi. Yolov3: An incremental im-provement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M.

Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 1, 2, 3, 4

[51] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 8

[52] 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*, 115:211–252, 2015. 5, 6

[53] Alberto Sabater, Luis Montesano, and Ana C. Murillo. Robust and efficient post-processing for video object detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10536–10542, 2020. 6

[54] Yuheng Shi, Naiyan Wang, and Xiaojie Guo. Yolov: Making still image object detectors great at video object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2254–2262, Jun. 2023. 2, 3, 4, 5, 6, 7, 8

[55] Guanxiong Sun, Yang Hua, Guosheng Hu, and Neil Robertson. Mamba: Multi-level aggregation via memory bank for video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2620–2627, 2021. 2, 3, 5, 6

[56] Guanxiong Sun, Chi Wang, Zhaoyu Zhang, Jiankang Deng, Stefanos Zafeiriou, and Yang Hua. Spatio-temporal prompting network for robust video feature extraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13587–13597, October 2023. 2, 5, 6

[57] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14454–14463, June 2021. 1

[58] 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 (CVPR)*, pages 5443–5452, June 2021. 2

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3, 5

[60] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7942–7951, 2019. 1

[61] Chi Wang, Yang Hua, Zheng Lu, Jian Gao, and Neil Robertson. Temporal meta-adaptor for video object detection. In *British Machine Vision Conference*, volume 2021, 2021. 2, 3

[62] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 6

[63] Han Wang, Jun Tang, Xiaodong Liu, Shanyan Guan, Rong Xie, and Li Song. Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 732–747, Cham, 2022. Springer Nature Switzerland. 2, 3

[64] 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)*, September 2018. 2, 3

[65] 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 (ICCV)*, October 2019. 1, 2, 3, 4, 5, 6, 7, 8

[66] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 8

[67] 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. 2, 3

[68] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook, 2021. 8

[69] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 1

[70] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 1–21, Cham, 2022. Springer Nature Switzerland. 8

[71] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: End-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7853–7869, 2023. 2, 3, 5, 6

[72] Tianfei Zhou, Fatih Porikli, David Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation, 2022. 8

[73] 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 (CVPR)*, June 2018. 2, 3

[74] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. *CoRR*, abs/2010.04159, 2020. 1, 3

[75] 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 (ICCV)*, Oct 2017. 1, 2

[76] 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 (CVPR)*, July 2017. 2