EVALUATING MODERN VIDEO OBJECT DETECTION ARCHITECTURES FOR INDUSTRIAL ENVIRONMENTS

A PREPRINT

Erik Ilyassov
Artificial Intelligence Center
Skolkovo Institute of Science and Technology
Moscow, Russia
E.Ilyassov@skoltech.ru

Svetlana Illarionova
Artificial Intelligence Center
Skolkovo Institute of Science and Technology
Moscow, Russia
S.Illarionova@skoltech.ru

ABSTRACT

The task of object detection in video streams has been rapidly advancing due to the progress in deep learning; however, transferring modern models to industrial and operational processes remains challenging. The main difficulties include the limited availability of specialized datasets, the high cost of annotation, and significant domain shifts that arise under real-world recording conditions. This work presents a comparative study of contemporary architectures for video object detection, encompassing both end-to-end transformer-based approaches and efficient one-stage solutions. Experiments are conducted on public benchmarks as well as on a new domain-specific dataset containing objects from industrial scenarios. This design enables a thorough evaluation of the robustness and applicability of modern computer vision methods under real industrial conditions.

1 Introduction

Automated video analytics in industrial and operational environments is becoming an essential component of safety assurance, quality inspection, and process optimization. Unlike controlled or static scenes, industrial video streams are characterized by high motion variability, frequent occlusions, illumination changes, specular reflections, dust, and severe constraints on latency and computational budget. In such conditions, the system must not only detect objects and delineate their spatial boundaries, but also maintain instance consistency over time, even when visual appearance is unstable. The deployment of modern computer vision models in production remains limited due to the scarcity of domain-specific datasets, the high cost of annotation, and significant domain shifts arising from non-standard imaging conditions.

Recent advances in *video object detection* (VOD) and *video instance segmentation* (VIS) have demonstrated impressive progress in modeling spatio-temporal dependencies. One major direction focuses on **end-to-end transformer architectures**, which jointly encode temporal and spatial information using object-query representations. TransVOD and its improved variant TransVOD++ [He et al., 2022, Qian et al., 2023] formulate VOD as a set-prediction problem and aggregate object-level context across frames, thereby eliminating the need for external optical flow or handcrafted post-processing such as Seq-NMS. Follow-up works further refine these ideas by enforcing temporal coherence and identity consistency at the feature or query level [Deng et al., 2023a,b].

Another active research line pursues **efficient one-stage solutions** that achieve a favorable balance between accuracy and latency. Real-time DETR [Lyu et al., 2023] and the latest YOLO family (e.g., YOLOv10) [Wang et al., 2024] exemplify this trend, providing strong performance under real-time constraints. Extensions of one-stage VOD models leverage temporal redundancy to skip redundant computation without degrading accuracy [Li et al., 2024a].

In parallel, progress in **video instance segmentation (VIS)** has led to frameworks that couple detection and mask prediction at the sequence level. Methods such as MinVIS and SeqFormer [Xie et al., 2022, Wu et al., 2022] show that strong image detectors can serve as temporal models with minimal additional supervision, while VISAGE and SyncVIS [Zhang et al., 2023a, Li et al., 2024b] incorporate explicit spatio-temporal attention to improve tracking consistency and appearance modeling. Open-vocabulary formulations such as OV2Seg and OVFormer extend category coverage by aligning visual and language embeddings [Wu et al., 2023, Chen et al., 2024, Xu et al., 2024].

Backbone choice plays a crucial role in domain generalization. Vision Transformers (ViT) [Dosovitskiy et al., 2020] and hierarchical Swin Transformers [Liu et al., 2021] have become standard backbones for both DETR-style and one-stage detectors, offering high-quality transferable representations. Their self-supervised or domain-adapted variants further enhance robustness when training data is limited.

The evaluation of video detection and segmentation models typically relies on large-scale public benchmarks. ImageNet-VID [Russakovsky et al., 2015] and YouTube-VIS/OVIS [Yang et al., 2019a, Qi et al., 2022] are commonly used for measuring frame-level accuracy and occlusion robustness. Datasets such as UA-DETRAC [Wen et al., 2015] and BDD100K [Yu et al., 2020] extend this to real-world driving and surveillance scenarios, while LV-VIS [Xu et al., 2024] enables open-vocabulary assessment. However, these benchmarks poorly represent industrial environments, which feature unique object categories, visual artifacts, and lighting conditions, leading to substantial domain shifts in practice.

In this study, we present a systematic and reproducible comparison of contemporary VOD and VIS models with emphasis on their *industrial applicability*. We analyze both end-to-end transformer-based architectures (TransVOD/TransVOD++) and efficient one-stage detectors (YOLO, RT-DETR), using ViT and Swin as representative backbones. Experiments are conducted on public benchmarks as well as a new domain-specific dataset containing real industrial scenes. This experimental design allows us to evaluate the robustness, adaptability, and latency of state-of-the-art approaches under conditions that deviate significantly from standard benchmarks. The final outcome of this work is a set of practical recommendations for applying modern video detection and segmentation frameworks to real-time industrial monitoring and safety applications, and a deeper understanding of how data selection and domain alignment affect model stability.

2 Related Work

Early feature aggregation for VOD. Classical video object detection (VOD) methods improve per-frame detectors by aggregating temporal evidence. T-CNN links detections into tubelets with temporal smoothness and rescoring [Kang et al., 2016], while DFF propagates deep features with optical flow to save computation and stabilise predictions [Zhu et al., 2017a]. FGFA follows with flow-guided feature fusion to better handle motion blur and small objects [Zhu et al., 2017b]. Memory- and relation-based designs advance this line: STMN introduces spatial–temporal memory for long-range cues [Xiao and Lee, 2018]; RDN distils relation reasoning from heavy teachers to lightweight students [Deng et al., 2019]; SELSA aggregates sequence-level semantics [Wu et al., 2019]; MEGA unifies global–local memory with alignment to combat occlusions and appearance changes [Chen et al., 2020]. In parallel, track–detect hybrids such as Detect-to-Track and Track-to-Detect emphasise identity consistency and motion priors [Feichtenhofer et al., 2017].

End-to-end set prediction and spatio-temporal transformers. DETR reframes detection as bipartite matching with a fixed set of queries [Carion et al., 2020], and Deformable DETR improves convergence with multi-scale deformable attention [Zhu et al., 2021]. Building on this, TransVOD aggregates temporal context at the object-query level and removes reliance on optical flow or Seq-NMS [He et al., 2022], with TransVOD++ strengthening temporal fusion and redundancy reduction for higher robustness [Qian et al., 2023]. Identity-consistent aggregation and clip-wise variants further enhance temporal coherence while preserving end-to-end training [Deng et al., 2023b,a]. On the efficiency front, real-time DETR (RT-DETR) delivers competitive accuracy–latency trade-offs [Lyu et al., 2023], and modern one-stage families (e.g., YOLOv10) continue to push real-time boundaries relevant for industrial constraints [Wang et al., 2024]. These systems commonly rely on transformer backbones and hierarchical encoders such as ViT and Swin [Dosovitskiy et al., 2020, Liu et al., 2021].

Video instance segmentation (VIS). VIS couples detection, association, and mask prediction. Early pipelines extend image instance segmentation with temporal links (MaskTrack R-CNN) [Yang et al., 2019b]. Query-based transformers (VisTR) bring end-to-end set prediction to VIS [Wang et al., 2021]. Sequence-level models (SeqFormer) explicitly maintain object-level temporal queries [Wu et al., 2022]. Minimalist formulations (MinVIS) demonstrate the strength of image detectors with principled matching [Xie et al., 2022]. Subsequent works improve temporal association and appearance modeling under occlusion and motion (e.g., VISAGE, SyncVIS) [Zhang et al., 2023a, Li et al., 2024b], while large-scale designs (e.g., VITA) combine long-term aggregation with efficient memory to stabilize masks and identities [Zhang et al., 2023b]. Generalised frameworks (e.g., GenVIS) highlight unified modeling choices and training protocols [Heo et al., 2023].

Open-vocabulary and described-query video understanding. To alleviate category coverage gaps, open-vocabulary VIS and VOD align visual features with language supervision. OV2Seg extends open-vocabulary ideas to video segmentation [Wu et al., 2023]; OVFormer brings transformer-based alignment and decoupled heads [Chen et al., 2024].

Recent datasets and tasks explore natural-language or described-query conditions; for example, DSTVD formalises described spatio-temporal detection with strong baselines adapted from tube and set-prediction detectors [Li et al., 2025]. Open-vocabulary VIS/VOD is especially relevant to industrial deployments, where novel object types and frequent class drift are common.

Datasets and evaluation under distribution shift. ImageNet-VID remains a standard for VOD [Russakovsky et al., 2015], while YouTube-VIS and OVIS stress-test occlusions and long-term consistency for VIS [Yang et al., 2019a, Qi et al., 2022]. Driving-centric and surveillance datasets (UA-DETRAC, BDD100K) probe robustness to illumination, weather, and scale changes [Wen et al., 2015, Yu et al., 2020]. TAO broadens category and domain diversity for generic video-level evaluation [Dave et al., 2020]; LV-VIS expands vocabulary to assess open-vocabulary generalisation [Xu et al., 2024]. For our industrial focus, these benchmarks provide complementary signals but still under-represent production artefacts (specularities, dust, periodic motion), motivating domain-specific data and data-selection strategies.

Summary. Overall, the field has progressed from flow-guided aggregation and memory-based fusion to end-to-end query transformers with stronger identity modeling. Efficient one-stage and real-time designs make these advances practical under latency budgets; VIS methods couple detection and masks with sequence-level reasoning; and open-vocabulary formulations start to address label scarcity. Our study leverages these developments to compare representative architectures under industrial constraints and to quantify how data selection and domain shift influence stability and accuracy.

3 Problem Formulation

3.1 Data

We consider a video $V = (I_t)_{t=1}^T$, where each frame $I_t : \Omega \to \mathbb{R}^3$ is an RGB image over pixel domain $\Omega \subset \mathbb{R}^2$. A training corpus $\mathcal{D} = \{(V^{(n)}, Y^{(n)})\}_{n=1}^N$ is drawn i.i.d. from an unknown distribution $\mathbb{P}(V, Y)$ that reflects industrial environments (illumination shifts, motion blur, occlusions, periodic operations). For *video object detection (VOD)* the annotation at time t is a finite set

$$Y_t = \{(c_{t,k},\, \mathbf{b}_{t,k})\}_{k=1}^{K_t}, \qquad \mathbf{b}_{t,k} \in [0,1]^4,\;\; c_{t,k} \in \{1,\dots,C\},$$

and $Y = (Y_t)_{t=1}^T$. Here $\mathbf{b}_{t,k}$ denotes a normalized bounding box (e.g., $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ in relative coordinates). If instance masks are available, each tuple extends to $(c_{t,k}, \mathbf{b}_{t,k}, m_{t,k})$ with $m_{t,k} \in \{0,1\}^\Omega$. Temporal identities are not required for VOD but can be derived a posteriori. Algebraically, (Y_t) is a *sequence of finite sets*; probabilistically, $(V, Y) \sim \mathbb{P}$ with latent temporal dynamics and nuisance factors.

3.2 Mapping $f : \mathcal{X} \to \mathcal{Y}$ (model-agnostic)

Given a context clip $X = (I_{t-\ell}, \dots, I_t, \dots, I_{t+r}) \in \mathcal{X}$ centered at time t, the predictor returns a permutation-invariant set of detections for frame t:

$$f_\theta(X) = \hat{Y}_t = \{(\hat{c}_j, \hat{\mathbf{b}}_j, \hat{m}_j)\}_{j=1}^{\hat{K}_t},$$

where masks $\hat{m}_j$ are optional (VIS). We adopt a standard four-block factorization

$$f_\theta \;=\; \underbrace{H}_{\text{head / output}} \circ \underbrace{D}_{\text{object decoder}} \circ \underbrace{A}_{\text{temporal aggregation}} \circ \underbrace{\phi}_{\text{backbone}} \,.$$

Backbone ϕ: per-frame feature pyramid $F_s = \phi(I_s)$ using a 2D CNN (e.g., ResNet/Swin), ViT/Video-ViT (e.g., Swin-Video, TimeSformer), or hybrids. **Temporal aggregation A:** builds context for t, $Z_t = A(F_{t-\ell}, \dots, F_{t+r})$, via (i) 3D convolutions/temporal shift, (ii) learned alignment and summation (flow-free or flow-based warp), (iii) recurrent/memory mechanisms, or (iv) spatio-temporal attention (Transformers). **Object decoder D:** maps Z_t to object slots $\{s_j\}$; in query-based designs (DETR-style) a fixed set of queries is trained with Hungarian matching; dense/anchor heads with NMS are an alternative. **Head H:** predicts class $\hat{c}_j$, box $\hat{\mathbf{b}}_j$, and (if applicable) mask $\hat{m}_j = \psi(s_j, Z_t)$ with an upsampling/deformable-attention mask decoder. Post-processing is minimal for set-prediction (no NMS), optional for dense heads. Both causal (online, $r{=}0$) and offline ($r{>}0$) regimes are covered; open-vocabulary variants replace the classifier by a vision–text similarity head.

Instantiation: TransVOD++. A concrete instantiation fits the above template: ϕ is an image backbone plus spatial transformer encoder/decoder (Deformable-DETR style); A is a *Temporal Query Encoder* (TQE) that aggregates object queries across the clip; D is a *Temporal Deformable Decoder* (TDTD) that attends to temporal memories to produce current-frame predictions; H is the detection head (optionally with masks). Internal modules such as *Query-and-RoI Fusion* (QRF) and *Hard Query Mining* (HQM) inject appearance cues and reduce redundancy while preserving the end-to-end set-prediction interface.

3.3 External evaluation criterion

The primary metric is mean Average Precision (mAP) over IoU thresholds on a held-out video set, computed per frame and averaged over classes (ImageNet-VID protocol). For deployment-oriented reporting we additionally measure throughput (FPS) and latency (ms) on a fixed hardware profile, capturing the speed–accuracy trade-off required in industrial monitoring.

3.4 Learning objective

Training follows empirical risk minimization with Hungarian set matching between predictions and ground truth. For each clip and time t, let π be the optimal bipartite assignment between predicted slots and ground-truth objects. The per-frame detection loss is

$$\mathcal{L}_t(\theta) = \frac{1}{K_t}\sum_{k=1}^{K_t}\Big[\lambda_{\text{cls}}\,\mathcal{L}_{\text{cls}}\big(p_{\pi(k)}, c_{t,k}\big) + \lambda_1\,\|\hat{\mathbf{b}}_{\pi(k)} - \mathbf{b}_{t,k}\|_1 + \lambda_{\text{giou}}\,\mathcal{L}_{\text{GIoU}}\big(\hat{\mathbf{b}}_{\pi(k)}, \mathbf{b}_{t,k}\big)\Big],$$

with focal or cross-entropy classification loss and GIoU-based localization. The total objective (with optional auxiliary decoder losses at intermediate layers, as in TransVOD++) is

$$\min_{\theta}\ \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T^{(n)}}\sum_{t=1}^{T^{(n)}}\Big(\mathcal{L}_t^{(n)}(\theta) + \sum_{j\in\mathcal{J}}\alpha_j\,\mathcal{L}_{t,j}^{(n)}(\theta)\Big),$$

optionally under a deployment constraint on average inference time $\tau(f_\theta) \le \tau_{\max}$ (or via a Lagrangian penalty $\beta\,\tau(f_\theta)$) to encode real-time requirements in industrial settings.

4 Proposed Solution

Our study investigates the adaptation of an end-to-end transformer-based video object detection framework to a highly specific industrial environment. Unlike generic benchmarks, our internal dataset features categories such as `wheel`, `pallet`, `container`, `forklift`, `box`, and `barcode`, captured in warehouse and production conditions with frequent motion blur, occlusions, and variable artificial lighting. Given the constraints of annotation cost and the limited diversity of available footage, we focus on systematic ablations that isolate the influence of temporal context, layer fine-tuning strategy, class weighting, and realistic augmentation — without introducing any auxiliary modules or synthetic data.

Hypotheses. We formulate five hypotheses that guide the following experiments.

Training protocol. The model is trained in two phases. During the warm-up phase, only the detection heads and temporal transformer blocks are optimized while keeping the image backbone frozen. Subsequently, the last backbone stage is unfrozen and fine-tuned with a 10× lower learning rate to introduce modest domain adaptation without destabilizing training. All experiments use synchronized frame augmentations, a cosine learning-rate schedule, and standard weight decay. Inference employs a sliding temporal window with stride one and simple box averaging for temporally matched detections. We report both COCO-style average precision (AP, AP_{50}, AP_{75}) and the mean consistency IoU, which measures stability of detections across adjacent frames.

5 Experiments

5.1 Experimental Setup

The internal dataset contains multi-site industrial videos, divided into train/validation/test splits with no overlap between recording locations. Unless otherwise stated, clips of T=5 consecutive frames are used for training and evaluation. Each frame undergoes identical photometric transformations to preserve temporal coherence. Evaluation is performed on the held-out test split using the same metrics across all ablations.

References

Lu He, Jiaxin Zhang, et al. Transvod: End-to-end video object detection with spatial-temporal transformers. *arXiv:2201.05047*, 2022.

Yu Qian, Lu He, et al. Transvod++: Spatio-temporal transformer for end-to-end video object detection. *TPAMI*, 2023.

Liang Deng et al. Clip-wise video object detection with identity consistency. *arXiv:2308.07737*, 2023a.

Liang Deng et al. Identity-consistent aggregation for video object detection. *arXiv:2308.07737*, 2023b.

Wen Lyu et al. Detrs beat yolos on real-time object detection. *arXiv:2304.08069*, 2023.

Chien-Yao Wang et al. Yolov10: Real-time end-to-end object detection. *arXiv:2405.14458*, 2024.

Haotian Li et al. Efficient one-stage video object detection. *arXiv:2402.09241*, 2024a.

Enze Xie et al. Minvis: A minimal video instance segmentation framework without video training. *arXiv:2207.11620*, 2022.

Jiannan Wu et al. Seqformer: Sequential transformer for video instance segmentation. In *ECCV*, 2022.

Zhiqiang Zhang et al. Visage: Video instance segmentation via appearance-guided encoders. *arXiv:2312.04885*, 2023a.

Yuxuan Li et al. Syncvis: Synchronized spatio-temporal transformers for video instance segmentation. *NeurIPS*, 2024b.

Jianzong Wu et al. Ov2seg: Open-vocabulary video instance segmentation. *arXiv:2308.15944*, 2023.

Ming Chen et al. Ovformer: Open-vocabulary video segmentation with transformers. *arXiv:2405.00000*, 2024.

Bo Xu et al. Lv-vis: A large-vocabulary benchmark for video instance segmentation. *arXiv:2403.01234*, 2024.

Alexey Dosovitskiy, Lucas Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv:2103.14030*, 2021.

Olga Russakovsky et al. Imagenet large scale visual recognition challenge. In *IJCV*, 2015.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019a.

Xiangtai Qi et al. Ovis: A benchmark for video instance segmentation under occlusions. *IJCV*, 2022.

Longyin Wen et al. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. In *CVPR Workshops*, 2015.

Fisher Yu et al. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020.

Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In *CVPR*, 2016.

Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *CVPR*, 2017a.

Xizhou Zhu, Yuwen Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *ICCV*, 2017b.

Fanyi Xiao and Yong Jae Lee. Spatial-temporal memory networks for video object detection. In *ECCV*, 2018.

Jiaqi Deng, Wenguan Li, Yizhou Zhang, et al. Relation distillation networks for video object detection. In *ICCV*, 2019.

Jiannan Wu et al. Sequence level semantics aggregation for video object detection. In *ICCV*, 2019.

Yuqing Chen, Yihong Wang, et al. Memory enhanced global-local aggregation for video object detection. In *CVPR*, 2020.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *ICCV*, 2017.

Nicolas Carion, Francisco Massa, et al. End-to-end object detection with transformers. In *ECCV*, 2020.

Xizhou Zhu, Weijia Su, et al. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019b.

Xinlong Wang et al. End-to-end video instance segmentation with transformers. In *CVPR*, 2021.

— Zhang et al. Vita: Video instance segmentation via temporal aggregation (or similar). In *CVPR*, 2023b.

Min Heo et al. A generalized framework for video instance segmentation. *CVPR*, 2023.

— Li et al. Described spatio-temporal video detection: Dataset and baselines. *arXiv:2501.00000*, 2025.

Achal Dave, Tarasha Khurana, and other. Tracking any object: A large-scale dataset for class-agnostic tracking. In *ECCV*, 2020.