Adapting Transformer-based Video Object Detection to Industrial Environments

Erik Ilyassov

Motivation

Problem. Most video detection models are trained on natural scenes (COCO, ImageNet VID). Industrial data differ drastically: warehouses, forklifts, boxes, tools, poor lighting, blur, occlusion. Our goal is to adapt **TransVOD++** to real industrial data.

Dataset.

- 19 classes (wheel, pallet, container, forklift, etc.)
- Multi-site warehouse videos, annotated manually

Temporal Component and Fine-Tuning Study

Main research question. How much does the *temporal component* contribute to recognition quality in industrial videos? Unlike public benchmarks (e.g., ImageNet VID, YouTube-VIS), there are **no industrial datasets** with temporal annotations — only isolated frames. Our goal is to analyze how temporal modeling affects performance when transferring to this domain.

Motivation. The pretrained TransVOD++ learns object motion patterns from natural scenes — animals, vehicles, people — which are very different from industrial motion (e.g., slow forklifts, repetitive pallet movement). We start from this model and replace only the detection head to match our 19 industrial categories, then compare baseline performance before and after fine-tuning.

Baseline results.

Model / Setting	Dataset	$\text{mAP}_{50:95}$	mAP_{50}	F1
TransVOD++ (Swin-Base, pretrained)	ImageNet VID	0.67	0.85	0.92
+ head replaced (before fine-tuning)	Industrial	0.11	0.21	0.34
+ fine-tuned on industrial data (ours)	Industrial	0.34	0.56	0.76

Interpretation. Replacing the classification head leads to a dramatic drop in accuracy - confirming a strong domain gap.