

SeaClips: Supplementary Material.

Franziska Denk^{1,2}

Christian Rankl¹

Shaban Almouahed¹

David Moser¹

Robert Sablatnig²

¹SEA.AI ²TU Wien

franziska.denk@sea.ai

A. Additional Material: Dataset

Additional Statistics. Table 1 provides a detailed breakdown of the number of bounding boxes per object category, per bounding box size, and per dataset split. As noted in the paper, it is not possible to create an independent split of videos containing *leisure vehicles* across the train, validation, and test sets. Consequently, the validation split contains no objects from the *leisure vehicle* category.

Datasheet. Ge *et al.* [4] propose the *Datasheet for Datasets* to properly document machine learning datasets. A datasheet for SeaClips can be found in a separate file in the supplementary material.

B. Additional Material: Experiments

This section provides further material on the experiments from the paper. While some details are already mentioned in the paper, they are included again for completeness.

B.1. Model Details

For all the models, the open source implementations are used [3, 10, 14–16]. The work of the other authors and the publication of their code is acknowledged gratefully.

YOLOX [4]. For the architecture, YOLOX-S, YOLOX-L and YOLOX-X are implemented as in their official implementation. The Intersection over Union (IoU) threshold for which bounding boxes are suppressed in Non-Maximum Suppression (NMS) is set to 0.4, and NMS is performed class-agnostically.

T-YOLOX-S [2, 9]. The implementation of T-YOLOX-S is based on YOLOX-S. It merges feature maps of the current frame and its reference frame after the first CSPLayer with a 3×3 convolution.

RT-DETR [17]. Similarly to YOLOX, RT-DETR architecture follows its official implementation. The number of queries is set to 300, and the number of feature maps is 3.

Deformable DETR [19]. While the Deformable DETR with ResNet-50 uses 4 feature maps, as in the official Deformable DETR implementation, the Deformable DETR with Swin-B backbone operates on a single feature map only, holding resized and summed up values from all Swin-B feature maps, as in the official TransVOD++ implementation. Deformable DETR with ResNet-50 backbone uses 300 queries, and with Swin-B backbone uses 100 queries.

TransVOD++ [18]. The base model is Deformable DETR with Swin-B using a single feature map. There are 100 object queries used, and the implementation is as in the official repository.

YOLOV++ [12, 13]. The boxes in the Feature Selection Module [13] are filtered according to a confidence threshold of 0.0001, and NMS is applied with a threshold of 0.5 IoU. NMS after the Feature Aggregation Module uses an IoU threshold of 0.4 and is performed in a class-agnostic way. Only classification scores are re-scored using the temporally aggregated video classification features, while keeping the objectness scores from the individual frames fixed.

B.2. Training Details

Gradients are clipped to a maximum norm of 0.1 during training for all models. Table 2 shows an overview of further hyperparameters used in the experiments.

Epochs. According to the implementation of YOLOX [4], for models trained with mosaic augmentation, a number of *no-augment epochs* is used at the end of training in which mosaic augmentation is disabled and L1-loss is used. As mentioned in the paper, only image-based YOLO models are trained with mosaic. Thus, for the remaining models, *no-augment epochs* is not applicable.

Pre-Training and Freezing Parameters. The *pre-trained modules* refers to the parts of the model that are pre-trained. The *pre-training dataset* refers to the dataset that is used to pre-train said modules. The *frozen parameters* specifies which parameters of a model are kept fixed during training, if applicable.

Category	Train			Validation			Test			Total
	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large	
Boat	3,625	2,048	588	1,725	0	0	4,159	854	0	12,999
Sailing vessel	15,015	8,929	2,607	487	412	504	10,509	2,966	301	41,730
Ship	905	517	385	481	601	0	549	307	828	4,573
Leisure vehicle	0	600	0	0	0	0	9,912	2,203	0	12,715
Animal	2,302	54	0	251	2	0	928	4	0	3,541
Marine marker	11,895	708	0	7,199	202	0	11,444	343	234	32,025
Object	15,175	189	0	1,835	0	0	4,416	0	0	21,615
Total	48,917	13,045	3,580	11,978	1,217	504	41,917	6,677	1,363	129,198

Table 1. **Detailed Object Counts in SeaClips.** Bounding box counts by dataset split, object category, and COCO [6] size category.

Optimizer and Learning Rate. Similarly, the implementation of YOLOX and YOLOV++ is followed for the warmcos learning rate scheduler of YOLOX, where a number of *warmup epochs* is defined in which the learning rate is increased from 0 to a *maximum learning rate*, from which it decays in a cosine schedule to a *minimum learning rate*. For the multi-step learning rate scheduler, the learning rate is multiplied by 0.1 at each scheduler step. Again, as in the open-source implementations, for Deformable DETR and TransVOD++, the learning rate for linear projection layers to obtain reference points and sampling offsets is the current learning rate multiplied by 0.1. The transformer models use a separate parameter group for the backbone, where the current learning rate is multiplied by a *learning rate multiplier*.

Losses. Losses are calculated as in the official implementations of the models. According to this, all transformer-based models are supervised with auxiliary losses, in addition to the losses calculated for the output at the last decoder layer.

B.3. Extended Results

The experiments in the paper are evaluated mainly with mAP@0.3. This section extends these results with additional material.

Statistical Analysis of Detection Performance. As stated in the paper, no improvement of a VID model over its base model in terms of mAP@0.3 is statistically significant. To arrive at this conclusion, a one-tailed Welch’s *t*-test with a significance level of 5% is performed on the three comparisons of VID models against their base models¹. To control the family-wise error rate across the three comparisons, a Bonferroni correction is applied, setting the significance threshold at $p < 0.01667$. Using this threshold, the observed performance gains for YOLOV++ (one-tailed $p = 0.085$) and TransVOD++ (one-tailed $p = 0.413$) are

¹ While YOLOV++ and TransVOD++ compared to the base model truly show a small improvement in mAP@0.3, which *could* be statistically significant, T-YOLOX-S is included in the statistical test only for completeness, as it showed no improvement over YOLOX-S at all, and thus, there is no improvement which could be statistically significant.

	YOLOX-S	YOLOX-L	YOLOX-X	T-YOLOX-S	YOLOV++
Ep.	15	15	15	7	7
No-augment ep.	5	5	5	-	-
Batch size	32	8	8	16	5
Scheduler	warmcos	warmcos	warmcos	warmcos	warmcos
Max. LR	2e-4	1e-4	1e-4	2e-4	1e-4
Min. LR	7e-6	2.5e-6	2.5e-6	5e-6	1e-6
Warmup ep.	2	2	2	2	2
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Weight decay	1e-4	1e-4	1e-4	1e-4	1e-4
Pre-trained	all	all	all	base model	base model
Frozen	-	-	-	-	base model*
Pre-train. data	I-VID [11]	I-VID [11]	I-VID [11]	SeaClips	SeaClips

(a) YOLO-based models. * excl. linear projection layers in its head

	Def. DETR (R50)	RT-DETR	Def. DETR (Swin-B)	TransVOD++
Ep.	15	15	15	7
Batch size	2	2	2	1
Scheduler	multi-step	multi-step	multi-step	multi-step
Max. LR	2e-4	1e-4	2e-5	2e-5
Min. LR	2e-6	1e-6	2e-7	2e-7
LR multiplier backbone	0.1	0.1	1	1
Warmup ep.	-	-	-	-
Optimizer	AdamW	AdamW	AdamW	AdamW
Weight decay	1e-4	1e-4	1e-4	1e-4
Pre-trained	all	all	all	base model
Frozen	-	-	-	base model
Pre-train. data	COCO [6]	COCO [6]	COCO [6]	SeaClips

(b) Transformer-based models.

Table 2. **Hyperparameters of the Experiments.** *Pre-trained* refers to the parts of the models that are pre-trained, *Frozen* refers to those that are kept frozen during training, *Pre-train. data* refers to the dataset that is used for pre-training. *Epochs* are abbreviated as *Ep.*, *Learning Rate* as *LR*, and ImageNet VID as I-VID.

not sufficient to reject the null hypothesis. Thus, the increases are not statistically significant.

Qualitative Examples. Additional examples of predictions on the test set from the models evaluated in the paper are shown in Figure 1. In Figure 1a, seven out of nine models fail to localize the *ship* with a correct bounding box shape and aspect ratio, as described in the paper. Eight out of nine

Method	mAP@0.5	mAP@0.5 by Size		
		Small	Medium	Large
<i>Image-based Detectors</i>				
YOLOX-S [4]	12.78 \pm 0.94	10.79	24.40	19.67
YOLOX-L [4]	17.57 \pm 1.82	16.64	21.62	33.71
YOLOX-X [4]	17.58 \pm 4.62	15.78	32.22	21.11
Def. DETR (R50) [19]	14.87 \pm 0.71	12.06	41.71	39.17
RT-DETR [17]	21.02 \pm 2.64	17.99	42.36	50.24
Def. DETR (Swin-B) [19]	<u>21.70</u> \pm 2.93	13.77	60.97	69.39
<i>Video-based Detectors</i>				
T-YOLOX [2, 9]	10.00 \pm 1.61	8.33	18.37	22.87
TransVOD++ [18]	23.50 \pm 3.33	14.22	54.32	61.28
YOLOV++ [13]	13.85 \pm 1.90	11.07	38.55	28.99

Table 3. **Overall Detection Performance with Increased IoU.** Results when an IoU threshold of 0.5 is used during evaluation. Object sizes are categorized according to the COCO protocol. Results are averaged over three runs and reported as mean \pm standard deviation in the mAP@0.5 column, and as mean in the other columns.

models also fail in categorizing the *ship*. Figure 1b shows a *marine marker*, where all YOLO-based models except for T-YOLOX-S fail to predict the correct class.

mAP at Higher IoU. Table 3 shows metrics calculated at an IoU of 0.5, *i.e.*, mAP@0.5. Comparing these metrics to the results given in the paper illustrates the impact of lowering IoU from 0.5 (as usually used according to the COCO protocol [6]) to 0.3 (as discussed in the paper).

Robustness to Environmental Conditions. To measure the performance of the models under different environmental conditions, a video-level annotation-conditioned evaluation could be performed. But, as the test data split is limited in size, these results would lack generalizability (*e.g.*, Is the performance of the models in the three rainy test videos determined by the rain itself or by the general characteristics of the scene, such as rare objects?). Instead, the evaluation is performed on two modified versions of the test set. The weather augmentations of the Python library Albumentations [1] are used to simulate different environmental conditions in the modified test sets. To simulate rain in one version, all images are modified using RandomRain with a probability of 1, a rain type of *drizzle*, a drop width of 10 pixels, a drop length of 20 pixels, and (100, 100, 100) as drop color. To simulate sun reflections in the other version, all images are modified using RandomSunFlare with a probability of 1, the *physics based* method², (0, 0, 1, 0.3) as the normalized region where the sun flare can occur, and a sun radius of 120 pixels. Table 4 shows the results. Rain conditions result in substantial reductions in mAP@0.3, particularly for small and medium-sized objects. On the contrary, sun flare con-

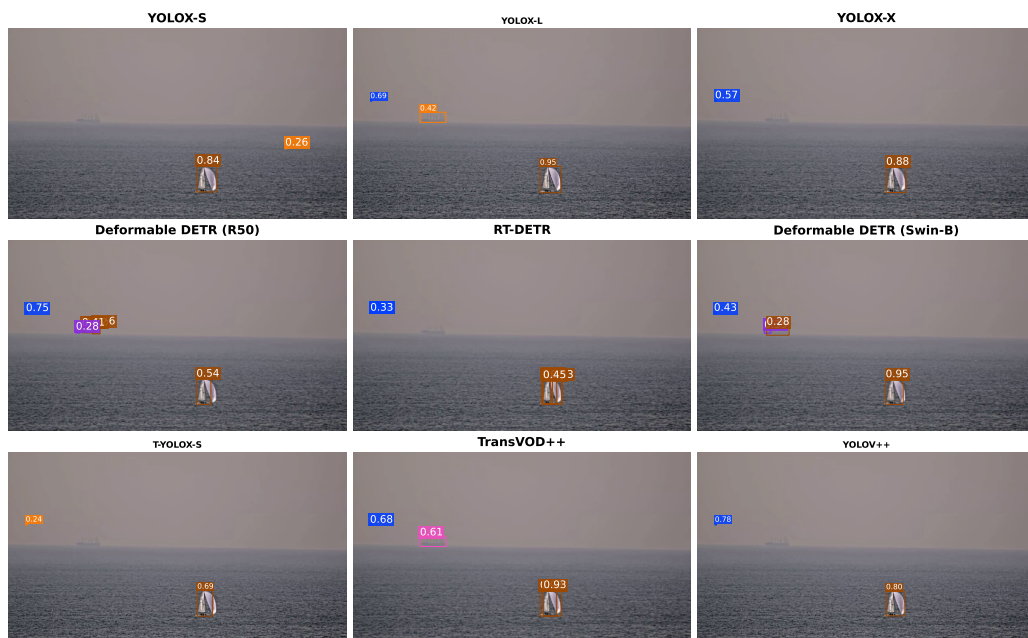
²Albumentations offers two methods to insert sun flare in images: *overlay* and *physics based*. The second method results in sun flares that look more realistic compared to sun flares created with the first method.

ditions reduce mAP@0.3 by only a few absolute percentage points. No weather transforms were included in training the models. Including them in the future would help to achieve models that are more robust against these (synthetic) weather conditions.

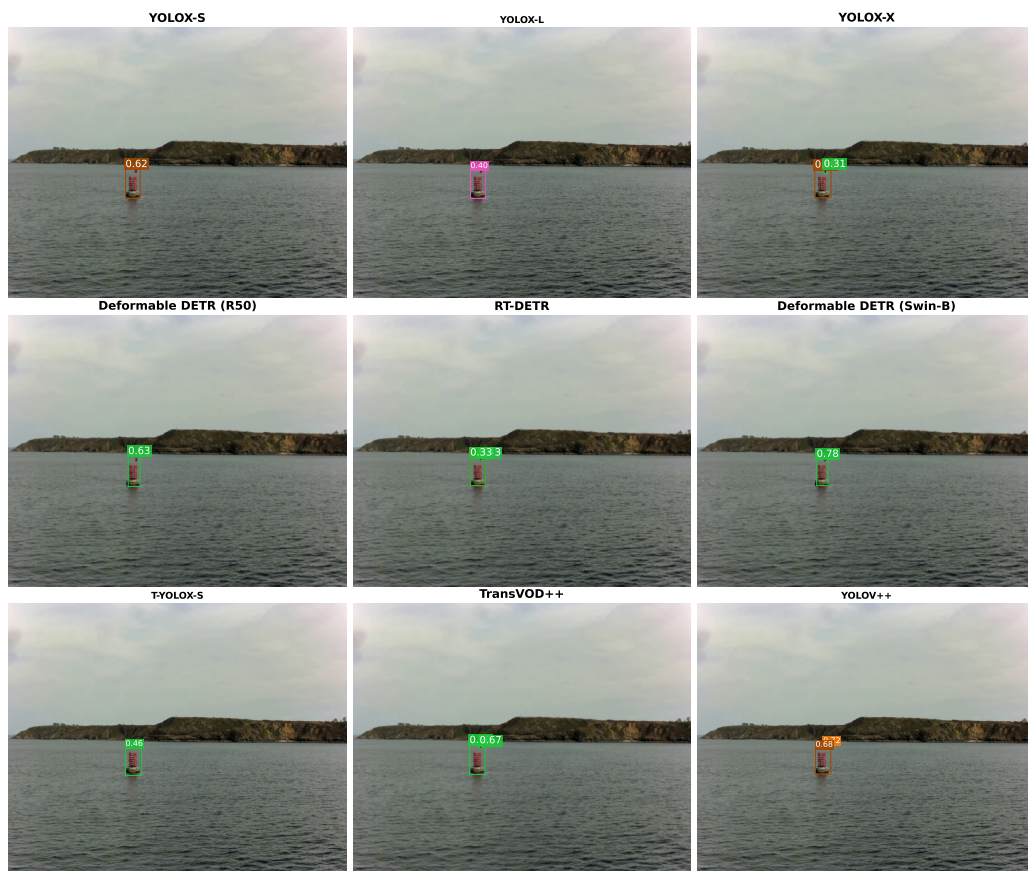
Cross-Dataset Transfer. To assess how well models trained on related datasets transfer to SeaClips, YOLOX-S and RT-DETR are trained on the Singapore Maritime Dataset (SMD) [8] and tested on the SeaClips test set. For these experiments, the training and validation sets of SMD, as proposed in [7], serve as the basis. As this validation set contains only three videos, it is augmented with two additional videos (MVI 1470 and MVI 1471) from SMD-Plus [5] for these experiments. SMD-Plus provides new annotations for previously unlabeled SMD videos. Since the object classes in SMD are different from those in SeaClips, Table 5 shows the class mapping that is used to align the predictions of SMD-trained models with SeaClips categories during evaluation. Although the mapping of *flying bird/plane* to *animal* is imperfect, the remaining mappings are semantically consistent. Table 6 shows the detection performance of the models trained in SMD and evaluated in SeaClips. As expected, the domain shift between the datasets results in inferior results compared to the models trained on SeaClips, as reported in the paper. Especially detection performance of small objects suffers from cross-dataset training. The exception to the degraded performance is the detection performance for large objects, which is improved when trained on SMD compared to when trained on SeaClips. These observations reflect the SMD dataset’s heavy domination by large bounding boxes of container ships and the limited presence of small objects. In contrast, SeaClips is dominated by small objects of other classes (*cf.*, Table 1). Table 7 shows the detection performance per class of models trained on SMD and evaluated on SeaClips. It validates the previous claims, as the detection performance of *ship* is increased when training with SMD, while the performance of other classes generally deteriorates.

C. Additional Material: Discussion

The results, presented in the main paper and extended in the supplementary material, show that maritime perception remains a challenging domain for computer vision. Improving robustness will require larger and more diverse datasets. Architectural innovations tailored to maritime conditions could also help improve the robustness and performance of computer vision in maritime domains in the future. Specifically, making domain-specific adaptations to the VID models could help to realize the potential of spatio-temporal modeling. SeaClips provides a foundation for this research by offering temporally dense annotations and realistic scenarios, paving the way for safer navigation at sea.



(a) *Sailing vessel close-by, ship visible at horizon in hazy conditions, and animal flying through the video.*



(b) *Marine marker.*

Figure 1. **Qualitative Examples.** Predictions on the test set. Best viewed zoomed in and in color.

■ Animal
 ■ Boat
 ■ Marine marker
 ■ Leisure vehicle
 ■ Object
 ■ Sailing vessel
 ■ Ship

Method	Rain					Sun Flare				
	mAP@0.3	mAP@0.3 by Size			mAP@0.3	mAP@0.3 by Size				
		Small	Medium	Large		Small	Medium	Large		
Image-based Detectors										
YOLOX-S [4]	9.46 -8.20	9.32 -6.61	13.17 -13.75	25.30 +4.16	17.47 -0.19	15.86 -0.07	26.36 -0.56	20.98 -0.16		
YOLOX-L [4]	19.13 -3.23	18.47 -2.61	21.58 -3.29	29.53 -8.94	21.87 -0.49	20.74 -0.34	24.34 -0.53	38.36 -0.11		
YOLOX-X [4]	16.87 -4.60	15.51 -4.17	29.16 -4.80	26.42 +2.00	20.96 -0.51	19.14 -0.54	36.51 +2.55	24.40 -0.02		
Def. DETR (R50) [19]	5.52 -17.34	4.19 -15.03	10.59 -34.85	26.38 -23.70	18.16 -4.70	15.28 -3.94	35.21 -10.23	42.37 -7.71		
RT-DETR [17]	9.23 -18.92	9.06 -15.92	21.03 -26.19	18.29 -33.94	27.68 -0.47	24.53 -0.45	42.90 -4.32	52.02 -0.21		
Def. DETR (Swin-B) [19]	23.71 -7.57	14.99 -7.94	55.76 -7.96	75.83 +1.77	30.11 -0.54	21.91 -1.02	62.72 -1.00	74.15 +0.09		
Video-based Detectors										
T-YOLOX [2, 9]	2.01 -15.01	2.33 -12.06	0.54 -19.58	8.10 -27.56	16.48 -4.60	14.17 -0.22	18.51 -1.61	34.59 -1.07		
TransVOD++ [18]	24.47 -7.48	16.15 -6.48	42.00 -15.01	67.91 +4.71	32.64 +0.69	22.76 +0.13	58.90 +1.89	62.30 -0.90		
YOLOV++ [13]	8.02 -12.34	5.87 -11.44	16.91 -26.45	26.73 -8.46	21.00 +0.64	16.46 -0.85	40.78 -2.58	35.18 -0.01		

Table 4. **Overall Detection Performance under Rain and Sun Flare Conditions.** Results when an IoU threshold of 0.3 is used during evaluation. Object sizes are categorized according to the COCO protocol. Results are averaged over three runs. Absolute differences to the metrics calculated on the non-modified test images (as given in the paper) are reported in subscripts.

SMD classes	SeaClips class
<i>Speed boat, Boat</i>	<i>Boat</i>
<i>Sail boat</i>	<i>Sailing vessel</i>
<i>Ferry, Vessel/Ship</i>	<i>Ship</i>
<i>Kayak</i>	<i>Leisure vehicle</i>
<i>Flying bird/plane</i>	<i>Animal</i>
<i>Buoy</i>	<i>Marine Marker</i>
<i>Swimming person, Other</i>	<i>Object</i>

Table 5. **Class Mapping: SMD to SeaClips.** Each row corresponds to a class mapping from SMD classes to a SeaClips class.

Method	Overall Performance		mAP _{ca} @0.3 by Size			mAP@0.3 by Size		
	mAP _{ca} @0.3	mAP@0.3	Small	Medium	Large	Small	Medium	Large
YOLOX-S [4]	13.16	6.56	10.03	40.54	53.73	0.72	14.27	36.48
RT-DETR [17]	8.80	9.43	6.63	32.28	24.43	4.82	38.04	55.20

Table 6. **Cross-Dataset Detection Performance.** Models are trained on SMD and evaluated on the SeaClips test set. Results are measured by mAP_{ca}@0.3 and mAP@0.3 for all samples of the test dataset, and split up by object sizes. Object sizes are categorized according to the COCO protocol.

Method	AP@0.3 by Object Class						
	Boat	Sailing vessel	Ship	Leisure vehicle	Animal	Marine marker	Object
YOLOX-S [4]	4.32	2.41	36.11	2.93	0.00	0.15	0.00
RT-DETR [17]	17.94	8.55	30.71	5.78	0.00	2.99	0.00

Table 7. **Cross-Dataset Detection Performance by Object Class.** Models are trained on SMD and evaluated on the SeaClips test set. Results are measured by AP@0.3 for each object category.

References

- [1] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2), 2020.
- [2] Christof W. Corsel, Michel van Lier, Leo Kampmeijer, Nicolas Boehrer, and Erwin M. Bakker. Exploiting Temporal Context for Tiny Object Detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE, 2023.
- [3] Deformable DETR repository. <https://github.com/fundamentalvision/Deformable-DETR>. Last accessed: 2025-09-11.
- [4] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO Series in 2021. *arXiv*, abs/2107.08430, 2021.
- [5] Jun-Hwa Kim, Namho Kim, Yong Woon Park, and Chee Sun Won. Object Detection and Classification Based on YOLO-V5 with Improved Maritime Dataset. *Journal of Marine Science and Engineering*, 10(3), 2022.

- [6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*. Springer International Publishing, 2014.
- [7] Sebastian Moosbauer, Daniel Konig, Jens Jakel, and Michael Teutsch. A Benchmark for Deep Learning Based Object Detection in Maritime Environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2019.
- [8] Dilip K. Prasad, Chandrashekar Krishna Prasath, Deepu Rajan, Lily Rachmawati, Eshan Rajabally, and Chai Quek. Object Detection in a Maritime Environment: Performance Evaluation of Background Subtraction Methods. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 20(5):1787–1802, 2019.
- [9] Yitong Quan, Benjamin Kiefer, Martin Meßmer, and Andreas Zell. Lightweight Multi-Frame Integration for Robust YOLO Object Detection in Videos. *arXiv*, abs/2506.20550, 2025.
- [10] RT-DETR repository. <https://github.com/lyuwenyu/RT-DETR>. Last accessed: 2025-09-11.
- [11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- [12] 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, 2023.
- [13] Yuheng Shi, Tong Zhang, and Xiaojie Guo. Practical Video Object Detection via Feature Selection and Aggregation. *ArXiv*, abs/2407.19650, 2024.
- [14] TransVOD++ repository. https://github.com/qianyuzyq/TransVOD_plusplus. Last accessed: 2025-09-11.
- [15] YOLOV++ repository. <https://github.com/YuHengsss/YOLOV>. Last accessed: 2025-09-11.
- [16] YOLOX repository. <https://github.com/Megvii-BaseDetection/YOLOX>. Last accessed: 2025-09-11.
- [17] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs Beat YOLOs on Real-time Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.
- [18] 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.
- [19] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations (ICLR)*, 2021.