VOVTrack: Exploring the Potentiality in Raw Videos for Open-Vocabulary Multi-Object Tracking
Supplementary Material

Zekun Qian[1,3]*, Ruize Han[2]*, Junhui Hou[3], Linqi Song[3], Wei Feng[1]†
[1]College of Intelligence and Computing, Tianjin University
[2]Shenzhen University of Advanced Technology
[3]City University of Hong Kong
{clarkqian, wfeng}@tju.edu.cn, hanruize@suat-sz.edu.cn, {jh.hou, linqi.song}@cityu.edu.hk

Appendix 1. OVMOT Problem Formulation

OVMOT requires the tracker to be capable of tracking objects from the open-vocabulary categories. We first present the problem formulation of this task from the training and testing stages.

At the training stage, the training data is $\{\mathbf{X}^{\text{train}}, \mathcal{A}^{\text{train}}\}$ that contains video sequences $\mathbf{X}^{\text{train}}$ and their respective annotations $\mathcal{A}^{\text{train}}$ of the objects. Given one frame in the video, each annotation $\alpha \in \mathcal{A}^{\text{train}}$ consists of a 2D bounding box $\mathbf{b} = [x, y, w, h]$, a unified ID d over the whole video, and a category label c, where (x, y) is the center pixel coordinates and (w, h) is the width and height of the box, the category belongs to the *base class* set, *i.e.*, $c \in \mathcal{C}^{\text{base}}$.

At the testing stage, the inputs consist of video sequences $\mathbf{X}^{\text{test}}$ and the set of all object classes $\mathcal{C} = \mathcal{C}^{\text{base}} \cup \mathcal{C}^{\text{novel}}$, where $\mathcal{C}^{\text{novel}}$ denotes the novel categories not appearing in the training set, *i.e.*, $\mathcal{C}^{\text{novel}} \cap \mathcal{C}^{\text{base}} = \emptyset$. OVMOT aims to obtain the trajectories of all objects in $\mathbf{X}^{\text{test}}$ belonging to classes $\mathcal{C}$. Each trajectory τ consists of a series of tracked objects τ_t at frame t, and each τ_t is composed of a 2D bounding box $\mathbf{b}$, and its object category c. Note that, during the testing stage, we need to evaluate not only the results on the base class $\mathcal{C}^{\text{base}}$, but also on the novel class $\mathcal{C}^{\text{novel}}$. The results on $\mathcal{C}^{\text{novel}}$ can validate the tracker's capability when facing objects from the open-vocabulary categories.

Appendix 2. Training Data Analysis

As discussed above, we use the training dataset in TAO for association module training. Next, we will analyze our experimental results from the perspective of the data quantity used for training.

TAO dataset. As shown in the first row of Table 1, we can see that the original TAO dataset has very few annotated frames, with only 18.1k frames, and limited box annotations of 54.7k. This is because the annotations in TAO were made at 1 FPS, resulting in a very limited number of supervised frames and available annotations for training a robust tracker.

As shown in the next row, in our self-supervised method, we use all the raw video frames without requiring any annotations. We can see that the usable frame quantity has increased by 30 times compared to the original training set (with annotations). Also, the quantity of available object bounding boxes for self-supervised training has reached 399.9k, which is 7.5 times as many as the original number of annotated ones. Moreover, by integrating the long-short-term sampling strategies, we can fully utilize all the long-short-term frames within the TAO raw videos through our self-supervised method, thereby achieving better results.

LVIS dataset. As shown in Table 1 in the main paper, the comparison methods QDTrack [3] and TETer [4] trained on the LVIS dataset with both base and novel classes, still yield poor results in TAO validation and test sets. This may be due to the imbalance in the data quantity of base and novel categories. Specifically, as seen in Table 1, although the LVIS dataset has a large number of frames and annotations for its base classes, the data for its novel classes is very limited, with the number of frames being $\frac{1}{66}$ and the number of annotations even less, at $\frac{1}{239}$.

Table 1. The number of frames and annotations can be used to train in LVIS, annotated TAO, TAO in our self-supervised paradigm.

Datasets	Frames		Annotations (detections)	
TAO (with GT)	18.1k		54.7k	
TAO (Raw videos)	534.1k		399.9k	
LVIS	base	novel	base	novel
	99.3k	1.5k	1264.9k	5.3k

*Equal contribution.
†Corresponding author.

Appendix 3. Module Complementarity Discussion

When designing the entire framework, we also consider the complementarity of the localization, classification, and association modules, enabling them to assist each other.

Improving classification via association. Following the baseline [5], we use the most frequently occurring category within a trajectory as the category classification result in that trajectory. This way, the classification results could be improved through better associations. Such relationship explains the reason that category clustering operations used in our self-supervised object association training strategy effectively increase the classification performance, as shown in Table 2 in the main paper.

Improving localization via association. Better tracking association also leads to more accurate detection evaluation in TETA metric. Specifically, TETA's localization accuracy (LocA) is not purely spatial-based but influenced by tracking consistency. When computing detection-to-GT matches, the matching score incorporates both spatial IoU and historical track association weights. This design allows consistently tracked detections to accumulate higher matching weights across frames, thus helping resolve ambiguous matches where multiple detections have similar spatial overlaps with a ground truth box. This makes improving LocA under TETA more challenging than conventional detection metrics, as it requires both accurate detection and consistent tracking. Nevertheless, our proposed tracking-related object-state-aware learning strategy effectively improves detection performance under this rigorous evaluation framework by maintaining high-quality tracking associations.

Similarly, better localization and classification results also help achieve improved association results, *e.g.*, the category clustering in our method for association learning could benefit from the better classification performance. This makes our entire framework a cohesive whole with multiple modules working collaboratively.

Appendix 4. More Experimental Analysis

Prompt-guided attention analysis. To demonstrate the effectiveness of prompt-guided attention in target state perception and illustrate the necessity of filtering out low-quality objects, we present additional examples of low prompt-guided attention in Figure 1. The targets shown in the figure exhibit severe occlusion, incompleteness, or poor recognizability, which aligns with our initial design considerations for the prompts. These damaged targets can lead to network training issues where learning ambiguous target features limits the network's Open-Vocabulary (OV) generalization capability. The proposed prompt-guided attention mechanism effectively suppresses this critical issue in OV settings, thereby significantly enhancing the perception of novel targets.

Table 2. Time complexity of different tracking methods

Method	Input shape	Para.	Model Size	FPS
QDTrack	(3,800,1334)	15.47M	298.6M	13.8
OVTrack	(3,800,1334)	16.52M	283.77M	1.8
Ours	(3,800,1334)	16.52M	283.77M	15.8

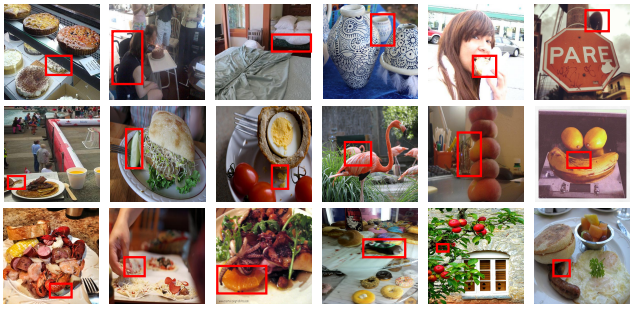

Figure 1. More visualization cases of the low prompt-guided attention targets.

Failure case analysis. We provide some failure cases in Figure 2. The first case illustrates a classification mistake due to significant occlusion. The second case shows the tracking errors caused by the distraction of object similarity and variability. We find that the OVMOT combined with the localization, classification, and tracking tasks presents a significant challenge, yet it holds large research potential.

Figure 2. Failure case illustration.

Model complexity analysis. We conducted a comparative analysis of inference speed (FPS) among different tracking methods, as shown in Table 2. Since the code for SLAck is not publicly available, we focused our comparison on two state-of-the-art methods with open-source implementations: QDTrack and OVTrack. Our experimental results demonstrate superior inference speed compared to these baseline methods. Notably, during the inference stage, we optimized the post-processing phase based on OVTrack by removing redundant NMS processing in the RPN and adjusting selection thresholds. These improvements significantly enhance the inference speed, achieving 15.8 FPS, which is approximately eight times faster than OVTrack (1.8 FPS) and outperforms QDTrack (13.8 FPS), while maintaining comparable tracking accuracy.

Analysis of parameter settings. We considered two representative parameter settings during the training and testing processes and conducted ablation experiments as

Table 3. Ablation studies on segment length and similarity threshold.

Parameter	Value	Base				Novel			
		TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
Segment Length	8	39.3	58.8	40.5	18.7	35.2	58.7	40.1	6.8
	16	39.5	58.8	40.7	18.9	35.2	58.7	40.0	6.7
	24	**39.6**	58.9	40.9	19.1	**35.3**	58.5	40.9	6.5
	30	39.2	58.4	40.3	18.7	34.9	58.4	39.3	6.9
Similarity Threshold	0.30	39.5	59.1	40.7	18.6	35.3	58.7	40.5	6.7
	0.35	**39.6**	58.9	40.9	19.1	**35.3**	58.5	40.9	6.5
	0.40	39.3	58.8	40.4	18.7	35.2	58.8	39.9	7.0
	0.45	38.9	58.4	39.8	18.6	34.5	58.2	38.2	7.2
	0.50	38.9	58.3	39.7	18.7	34.5	58.3	38.1	7.1

shown in Table 3. During training, we examine the impact of video segment length (in Section 3.3) used for self-supervised training on the final results. We can see that our method demonstrates good robustness to segment length, performing well across lengths from 8 to 30, with the best results at a length of 24 (used in the experiments). In the inference process, we evaluate the similarity score threshold (in Section 3.3) used in the association sub-task. We observe that the Base AssoA maintains stable performance around 40 within the range of 0.3 to 0.5. Similarly, the Novel AssoA remains steady at approximately 40 between 0.3 and 0.4, and slightly decreases to around 38 when the threshold exceeds 0.4. Both conditions demonstrate minimal fluctuation throughout their respective ranges. This shows that our algorithm is also not sensitive to the tracker's similarity threshold setting.

Analysis of method generalization. To evaluate the generalizability of our proposed self-supervised method for OVMOT, we conduct cross-domain (cross-dataset) self-supervised training and subsequently still validate the performance on the TAO validation set. Specifically, we replaced the unannotated TAO dataset with unlabeled videos from the LV-VIS dataset [6] during training. As shown in Table 4, despite a slight performance degradation in the cross-domain scenario, the performance loss is negligible, and the method maintains state-of-the-art (SOTA) performance. This experimental result demonstrates the *remarkable generalizability* of our proposed approach to the raw videos for training, highlighting its robust transfer learning capabilities across different video datasets.

Table 4. Cross-domain transfer performance of proposed self-supervised method.

Method	Base				Novel			
	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
Trained by LV-VIS [6]	39.2	58.5	40.3	18.7	34.8	58.5	40.1	5.7
Trained by TAO (Ours)	39.6	58.9	40.9	19.1	35.3	58.5	40.9	6.5

Appendix 5. More Supplementary Details

Details during training. As mentioned in the main paper regarding the experimental procedure, compared to using the existing Open-Vocabulary Detection (OVD) method [2] directly for localization and classification in OVTrack [5], we train the OVD process using the base classes of the LVIS dataset and incorporate tracking-related states into the training process (Section 3.2). This significantly enhanced the localization and classification results in open-vocabulary object tracking.

Additionally, in the training of the association module, different from our baseline method [5] using the generated image pairs constructed by LVIS, we further introduce a self-supervised method for object similarity learning (Section 3.3). Specifically, we utilize all the video frames in the TAO [1] training dataset for self-supervised training, which makes full use of the consistency among the objects in a video sequence and greatly improves the association task results.

Short-long-interval sampling strategy. We consider the interval splitting of $\mathcal{T}_c$ in Eq. (3). As shown in Figure 3, we split the original videos into several segments of length L and randomly sample the shorter sub-segments with various lengths from each segment. These short-term sub-segments are then concatenated to form the training sequence. Such training sequences include long-short-term intervals. Specifically, we select the adjacent frames from the same sub-segment, which allows the association head to learn the consistency objectives under minor object differences. We also select the long-interval video frames from different sub-segments, which allows the association head to learn the similarity and variation of objects under large differences.

Metrics. First, the localization accuracy (LocA) is determined through the alignment of all labeled boxes α with the predicted boxes of $\mathcal{T}$: $\text{LocA} = \frac{|\text{TPL}|}{|\text{TPL}|+|\text{FPL}|+|\text{FNL}|}$. Next, classification accuracy (ClsA) is calculated using all accurately localized TPL instances, by comparing the predicted semantic classes with the corresponding ground truth classes $\text{ClsA} = \frac{|\text{TPC}|}{|\text{TPC}|+|\text{FPC}|+|\text{FNC}|}$. Finally, association accuracy (AssocA) is determined using a comparable approach, by matching the identities of associated ground truth instances with accurately localized predictions $\text{AssocA} = \frac{1}{|\text{TPL}|}\sum_{b\in\text{TPL}}\frac{|\text{TPA}(b)|}{|\text{TPA}(b)|+|\text{FPA}(b)|+|\text{FNA}(b)|}$. The TETA score is computed as the mean value of the above three scores $\text{TETA} = \frac{\text{LocA}+\text{ClsA}+\text{AssocA}}{3}$.

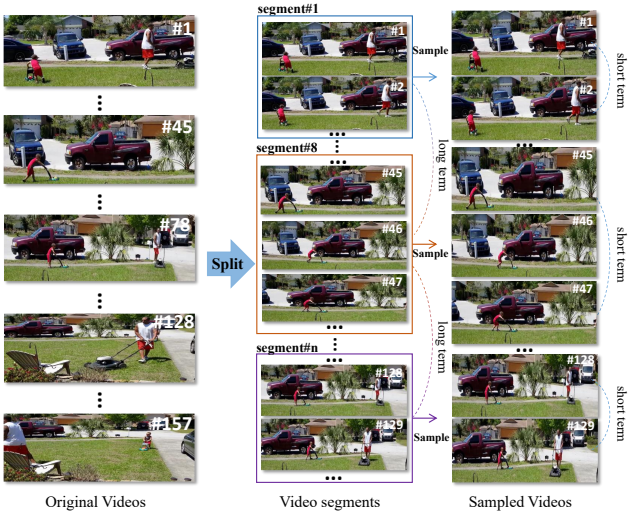

Figure 3. An illustration of interval sampling strategy.

References

[1] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In *European Conference on Computer Vision*, pages 436–454, 2020. 3

[2] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 3

[3] Tobias Fischer, Thomas E Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. QDTrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[4] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E Huang, and Fisher Yu. Tracking every thing in the wild. In *European Conference on Computer Vision*, pages 498–515, 2022. 1

[5] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5567–5577, 2023. 2, 3

[6] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 4057–4066, 2023. 3