TMT-VIS: Taxonomy-aware Multi-dataset Joint Training for Video Instance Segmentation

Anonymous Author(s)
Affiliation
Address
email

This supplementary material provides more details about the proposed TMT-VIS, further details of VIS datasets, more qualitative visual comparisons, and the codebase of our implementation. The content is organized as follows:

- More details of multiple VIS datasets.
- More ablation study experiments of the TMT-VIS.
- The qualitative visual comparisons between popular VIS method Mask2Former-VIS and TMT-VIS.
- The codebase is contained in the file 'TMT-VIS.zip'.

1 Dataset Details

Here, we provide a detailed overview of various VIS datasets in Table 1. Our extensive experimental evaluations are conducted on four challenging benchmarks, namely YouTube-VIS 2019 and 2021 [7], OVIS [4], and UVO [5]. YouTube-VIS 2019 [7] was the first large-scale dataset designed for video instance segmentation, comprising 2.9K videos averaging 4.61s in duration and 27.4 frames in validation videos. YouTube-VIS 2021 [7] poses a greater challenge with longer and more complex trajectory videos, averaging 39.7 frames in validation videos. The OVIS [4] dataset is another challenging VIS dataset with 25 object categories, focusing on complex scenes with significant object occlusions. Despite containing only 607 training videos, OVIS's videos last an average of 12.77s. Lastly, the UVO [5] dataset consists of 1.2K Kinetics-400 [3] videos, densely annotated at 30fps, featuring 81 object categories, including an extra "other" category for non-COCO instances. It provides exhaustive segmentation masks for all object instances in the 503 videos.

Among all categories in Youtube-VIS 2019, OVIS, and UVO, there are overlapping categories between each dataset, and there are also different categories that share similar semantics. The detailed overlapping categories are marked in Table 2. Overall, Youtube-VIS 2021 and OVIS share a more similar taxonomy space with Youtube-VIS 2019 than UVO, with a common category set of 34 out of 40 for Youtube-VIS 2021 and 22 out of 25 for OVIS. Typically, when the taxonomy spaces of datasets are similar, training them jointly will have smaller dataset biases, which leads to a better result in performance. The characteristics of these datasets align with the improvement in performance when validating the joint-training models on various datasets: the increase is more significant on Youtube-VIS 2021 and OVIS than on UVO. Further details of some specific categories can be found in Table 4.

2 Additional Ablation Studies

In this section, we provide more experiments on our proposed methods, we will discuss the generality and zero-shot properties of our training approach, and we will provide further details of the performance change in different taxonomies.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table 1: Key statistics of popular VIS datasets. Note that in UVO, the majority of the videos are for Video Object Segmentation, and only 503 videos are annotated for the VIS task. 'YTVIS' is the acronym of 'Youtube-VIS'.

	YTVIS19	YTVIS21	OVIS	UVO
Videos	2883	3859	901	11228
Categories	40	40	25	81
Instances	4883	8171	5223	104898
Masks	131K	232K	296K	593K
Masks per Frame	1.7	2.0	4.7	12.3
Object per Video	1.6	2.1	5.8	9.3

Generality property. The proposed new taxonomy-aware training strategy is an effective and general strategy that can be adapted into various DETR-based approaches (both online & offline) and into various datasets. As in Table 3, when adding our TCM&TIM strategy to the popular VIS architecture Mask2Former-VIS [1], VITA [2], and IDOL [6], we harvest performance gains on all of the three challenging benchmarks. In OVIS, the increase can be up to 6.3 AP for Mask2Former-VIS. As for the popular IDOL, our strategy can also bring about an increase of 2.8 AP in performance. This demonstrates that the proposed taxonomy-aware module can be treated as a plug-and-play design that can be used in various DETR-based VIS methods (both online & offline) across different scenarios and all popular VIS benchmarks.

Per-category performance. In Table 4, we present the comparison in performance between Mask2Former-VIS and our TMT-VIS on several specific taxonomies. When datasets other than YTVIS have no such taxonomy, training Mask2Former-VIS on multiple datasets will end up decreasing the performance of such category, as shown in 'Duck' case from Table 4. When applying our approach, we can obtain a performance improvement due to that taxonomy information of 'Duck' is compiled and injected to the instance queries. In other taxonomies, such as 'Person' which appears across all VIS datasets, the improvements are also significant.

Zero-shot property. Further, our TMT-VIS can also perform well on zero-shot learning. We conducted the experiments on Youtube-VIS 2019 [7], OVIS [4], and UVO [5] benchmarks. As exhibited in Table 5, TMT-VIS can be utilized to transfer the knowledge of other VIS datasets to another dataset with a significant increase of 4.5 AP and 3.8 AP respectively. The extra taxonomy information provided by our newly designed TCM & TIM improve the model's performance when dealing with unfamiliar taxonomies.

3 Visualization

In the visualization comparisons between Mask2Former-VIS and our model, we select some cases under different scenarios, which include setting with multiple similar instances, setting with fast-moving objects, and setting with different poses of instance.

In Fig. 1, we demonstrate videos with multiple similar instances, and TMT-VIS can both segment and track them more accurately than Mask2Former-VIS. The swimmers or the cyclists are all instances that belong to 'person' category, and TMT-VIS shows better segmentation and tracking. In Fig. 2, we present example videos which have quick movements in camera's perspective and have instances with different poses. In the top two rows, the sedan and the truck have similar appearances, and our model can distinguish and segment them with higher confidence (90% over 70%). In the last two rows, our model successfully segments the person in different poses, while Mask2Former-VIS fails to segment this person's arm in the first frame. However, the first two rows of Fig. 2 also show that our model still have the problem of segmenting instances with heavy occlusions. This suggests that simply combining taxonomy information is insufficient of solving severely occluded scenes, and that more information should be aggregated to instance queries to make the model more robust in segmenting instances.

Table 2: Overlapping categories of multiple VIS datasets with Youtube-VIS 2019 dataset.'YTVIS' is the acronym of 'Youtube-VIS'. As demonstrated in the table, YTVIS2021 and OVIS have a more similar taxonomy space, with 34 and 22 overlapping categories respectively.

YTVIS19_40_categories	**YTVIS21_40_categories**	**OVIS_25_categories**	**UVO_81_categories**
Overlapping Categories	34	22	19
Person	✓	✓	✓
Giant_panda	✓	✓	
Lizard	✓	✓	
Parrot	✓	✓	
Skateboard	✓		✓
Sedan		✓	✓
Ape			
Dog	✓	✓	✓
Snake	✓		
Monkey	✓	✓	
Hand			
Rabbit	✓	✓	
Duck	✓		
Cat	✓	✓	✓
Cow	✓	✓	✓
Fish	✓	✓	
Train	✓		✓
Horse	✓	✓	✓
Turtle	✓	✓	
Bear	✓	✓	✓
Motorbike	✓	✓	✓
Giraffe	✓	✓	✓
Leopard	✓		
Fox	✓		
Deer	✓		
Owl			
Surfboard			✓
Airplane	✓	✓	✓
Truck	✓	✓	✓
Zebra	✓	✓	✓
Tiger	✓	✓	
Elephant	✓	✓	✓
Snowboard	✓		✓
Boat	✓	✓	✓
Shark	✓		
Mouse	✓		
Frog	✓		
Eagle			
Earless_seal	✓		
Tennis_racket	✓		✓

Table 3: Ablation study on the generality property of TCM/TIM design with ResNet-50 backbone on multiple datasets.

Datasets	Method	AP	AP_{50}	AP_{75}
YouTube-VIS 2019	Mask2Former-VIS	46.4	68.0	50.0
	+ TCM/TIM	49.7 (↑ 3.3)	73.4 (↑ 5.4)	53.9 (↑ 3.9)
	VITA	49.8	72.6	54.5
	+ TCM/TIM	52.6 (↑ 2.8)	74.4 (↑ 1.8)	57.6 (↑ 3.1)
	IDOL	49.5	74.0	52.9
	+ TCM/TIM	51.4 (↑ 1.9)	74.9 (↑ 0.9)	55.0 (↑ 2.1)
YouTube-VIS 2021	Mask2Former-VIS	40.6	60.9	41.8
	+ TCM/TIM	44.9 (↑ 4.3)	66.1 (↑ 5.2)	48.5 (↑ 6.7)
	VITA	45.7	67.4	49.5
	+ TCM/TIM	48.3 (↑ 2.6)	69.8 (↑ 2.4)	50.8 (↑ 1.3)
	IDOL	43.9	68.0	49.6
	+ TCM/TIM	45.8 (↑ 1.9)	69.2 (↑ 1.2)	50.9 (↑ 1.3)
OVIS	Mask2Former-VIS	16.5	36.5	14.6
	+ TCM/TIM	22.8 (↑ 6.3)	43.6 (↑ 7.1)	21.7 (↑ 7.1)
	VITA	19.6	41.2	17.4
	+ TCM/TIM	25.1 (↑ 5.5)	45.9 (↑ 4.7)	23.8 (↑ 6.4)
	IDOL	30.2	51.3	30.0
	+ TCM/TIM	33.0 (↑ 2.8)	55.7 (↑ 4.4)	33.2 (↑ 3.2)

Table 4: Comparisons between per-category performance of Mask2Former-VIS and TMT-VIS. 'MDT' refers to 'Multiple Datasets Training', indicating whether the approach is trained on YTVIS, OVIS, and UVO. 'In Corresponding Dataset' is used to demonstrate whether the category is contained in corresponding dataset.

Categories	Methods	In Corresponding Dataset			MDT	Test Set	AP
		YTVIS	OVIS	UVO			
Person	Mask2Former-VIS	✓				YTVIS	57.2
	TMT-VIS	✓				YTVIS	57.9 (↑ 0.7)
	Mask2Former-VIS	✓	✓	✓	✓	YTVIS	59.3
	TMT-VIS	✓	✓	✓	✓	YTVIS	60.7 (↑ 1.4)
Duck	Mask2Former-VIS	✓				YTVIS	41.6
	TMT-VIS	✓				YTVIS	42.4 (↑ 0.8)
	Mask2Former-VIS	✓			✓	YTVIS	38.3
	TMT-VIS	✓			✓	YTVIS	43.9 (↑ 5.6)
Monkey	Mask2Former-VIS	✓				YTVIS	24.7
	TMT-VIS	✓				YTVIS	26.4 (↑ 1.7)
	Mask2Former-VIS	✓	✓		✓	YTVIS	25.6
	TMT-VIS	✓	✓		✓	YTVIS	29.1 (↑ 3.5)
Snowboard	Mask2Former-VIS	✓				YTVIS	8.9
	TMT-VIS	✓				YTVIS	11.8 (↑ 2.9)
	Mask2Former-VIS	✓		✓	✓	YTVIS	10.0
	TMT-VIS	✓		✓	✓	YTVIS	14.5 (↑ 4.5)

Table 5: Zero-shot Performance of TMT-VIS with ResNet-50 backbone. The results demonstrate the zero-shot ability of our proposed method. YouTube-VIS 2019 is abbreviated as 'YTVIS'.

Method	Train Set			Test Set	AP	AP_{50}	AP_{75}
	YTVIS	OVIS	UVO				
Mask2Former-VIS		✓	✓	YTVIS	7.1	11.4	8.3
TMT-VIS		✓	✓	YTVIS	11.6 (↑ 4.5)	17.2 (↑ 5.8)	15.0 (↑ 6.7)
Mask2Former-VIS	✓		✓	OVIS	3.7	9.8	5.2
TMT-VIS	✓		✓	OVIS	7.5 (↑ 3.8)	14.1 (↑ 4.3)	8.5 (↑ 3.3)

Figure 1: Visual comparison of our model with Mask2Former-VIS (abbreviated as 'M2F-VIS'). Our TMT-VIS shows better precision in segmenting and tracking small instances with the same taxonomy (The swimmers or the cyclists are all in 'person' category, and TMT-VIS shows better performance).

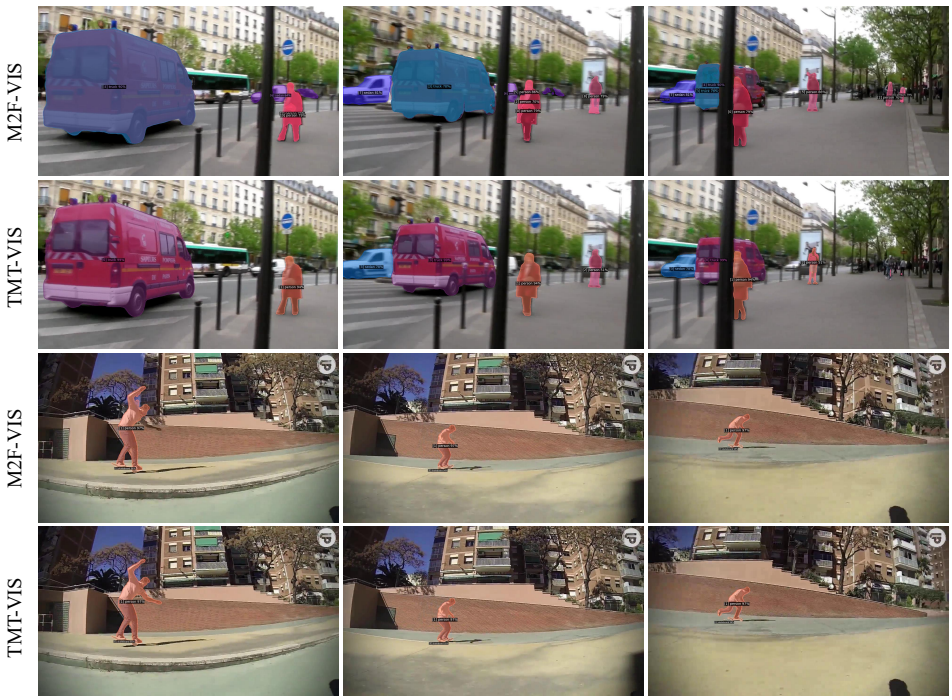

Figure 2: Visual comparison of our model with Mask2Former-VIS (abbreviated as 'M2F-VIS'). Our TMT-VIS shows better precision in segmenting and tracking instances with quick movements or with different poses. In the top two rows, the sedan and the truck have similar appearances, and our model can classify them with higher confidence. In the last two rows, our model successfully segments the person in different poses, while M2F-VIS fails to segment this person's arm in the first frame.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2

[2] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. In *NeurIPS*, 2022. 2

[3] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv:1705.06950*, 2017. 1

[4] 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. *IJCV*, 2022. 1, 2

[5] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. 1, 2

[6] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, 2022. 2

[7] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 1, 2