Supplementary Materials to 'MDQE: Mining Discriminative Query Embeddings to Segment Occluded Instances on Challenging Videos'

Minghan Li Shuai Li Wangmeng Xiang Lei Zhang*
Department of Computing, Hong Kong Polytechnic University
liminghan0330@gmail.com, {csshuaili, cswxiang, cslzhang}@comp.polyu.edu.hk

In this supplementary file, we provide the following materials:

- Visualization of initial query locations selected by our query initialization (cf. Section 4.2 in the main paper);
- Quantitative results with ResNet50 and ResNet101 on YouTube-VIS 2019 (cf. Section 4.3 in the main paper);
- Visualization of segmented instance segmentation on challenging videos (cf. Section 4.3 in the main paper).

A. Visualization of initial query locations

Fig. 1 visualizes the initial query locations selected by our query initialization method on challenging videos of OVIS valid set. We see that frame-level queries are well associated in both spatial and temporal directions, even in crowded scenes.

B. Quantitative results with ResNet50 and ResNet101 on YouTube-VIS 2019 valid set

The quantitative results on YouTube-VIS 2019 valid set are reported in Tab. 1. We see that the newly developed VIS methods utilizing transformer-based prediction heads can significantly improve the performance to 49.8% mask AP with ResNet50 backbone and 51.9% mask AP with ResNet101 backbone. Our proposed MDQE employs 5-frame clips as input and a deformable-attention decoder, reaching 47.3% and 47.9% mask AP with ResNet50 and ResNet101 backbones, respectively. The YouTube-VIS 2019 valid set contains mostly short videos. Therefore, methods like SeqFormer [11] and VITA [4], which take the video-in video-out offline inference, can incorporate the temporal information to distinguish the objects that have obvious differences in feature space. This is the reason why they achieve leading mask AP scores on YouTube-VIS 2019 valid set. However, for objects that have similar-looking appearance or heavy occlusions, methods like SeqFormer [11] and VITA [4] cannot extract discriminative features to distinguish them accurately, thereby resulting in unsatisfactory results on OVIS data set (cf. Section 4.3 in the main paper).

C. Visualization of segmented instance masks

Figs. 2 - 3 show the instance masks obtained by the recently proposed top-performing methods with ResNet50 backbone, including IDOL [12] (ECCV 2022), MinVIS [5] (NIPS 2022), VITA [4] (NIPS 2022) and our MDQE.

YouTube-VIS 2021 valid set. Fig. 2 shows the visual comparisons on YouTube-VIS 2021 valid set. We can see that all recent top-performing methods can well process the simple videos on YouTube-VIS 2021 valid set.

OVIS valid set. Fig. 3 and Fig. 4 compare the visual results of instance segmentation on OVIS valid set. Compared with IDOL [12] and MinVIS [5] with per-frame input, our MDQE with per-clip input can exploit richer spatio-temporal features of objects, thereby segmenting occluded instances better. On the other hand, MDQE with discriminate query embeddings can track instances with complex trajectories more accurately in challenging videos, such as cross-over objects and heavily occluded objects in crowded scenes.

*Corresponding author.

Figure 1. Visualization of object query locations by our proposed grid-guided query selection and inter-frame query association method, where the highlight dots with the same color represent the initialized locations of object queries on the same object.

Type	Method	ResNet 50					ResNet 101				
		AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Per-frame	MaskTrack [14]	30.3	51.1	32.6	31.0	35.5	31.8	53.0	33.6	33.2	37.6
	SipMask [3]	32.5	53.0	33.3	33.5	38.9	35.0	56.1	35.2	36.0	41.2
	STMask [7]	33.5	52.1	36.9	31.1	39.2	36.8	56.8	38.0	34.8	41.8
	SG-Net [9]	34.8	56.1	36.8	35.8	40.8	36.3	57.1	39.6	35.9	43.0
	CrossVIS [15]	34.8	54.6	37.9	34.0	39.0	36.6	57.3	39.7	36.0	42.0
	IDOL [12]	49.5	74.0	52.9	47.7	58.7	50.1	73.1	56.1	47.0	57.9
	MinVIS [5]	47.4	69.0	52.1	45.7	55.7	-	-	-	-	-
Per-clip	STEm-Seg [1]	30.6	50.7	33.5	31.6	37.1	34.6	55.8	37.9	34.4	41.6
	MaskProp [2]	40.0	-	42.9	-	-	42.5	-	45.6	-	-
	Propose-Reduce [8]	40.4	63.0	43.8	41.1	49.7	43.8	65.5	47.4	43.0	53.2
	VisTR [10]	35.6	56.8	37.0	35.2	40.2	40.1	64.0	45.0	38.3	44.9
	EfficentVIS [13]	37.9	59.7	43.0	40.3	46.6	39.8	61.8	44.7	42.1	49.8
	IFC [6]	41.0	62.1	45.4	43.5	52.7	42.6	66.6	46.3	43.5	51.4
	SeqFromer [11]	47.4	69.8	51.8	45.5	54.8	49.0	71.1	55.7	46.8	56.9
	VITA [4]	49.8	72.6	54.5	49.4	61.0	51.9	75.4	57.0	49.6	59.1
	MDQE (ours)	47.3	66.9	53.1	42.9	52.9	47.9	70.3	53.8	43.2	53.1

Table 1. Quantitative performance comparison of methods with ResNet50 and ResNet101 on YouTube-VIS 2019 valid set.

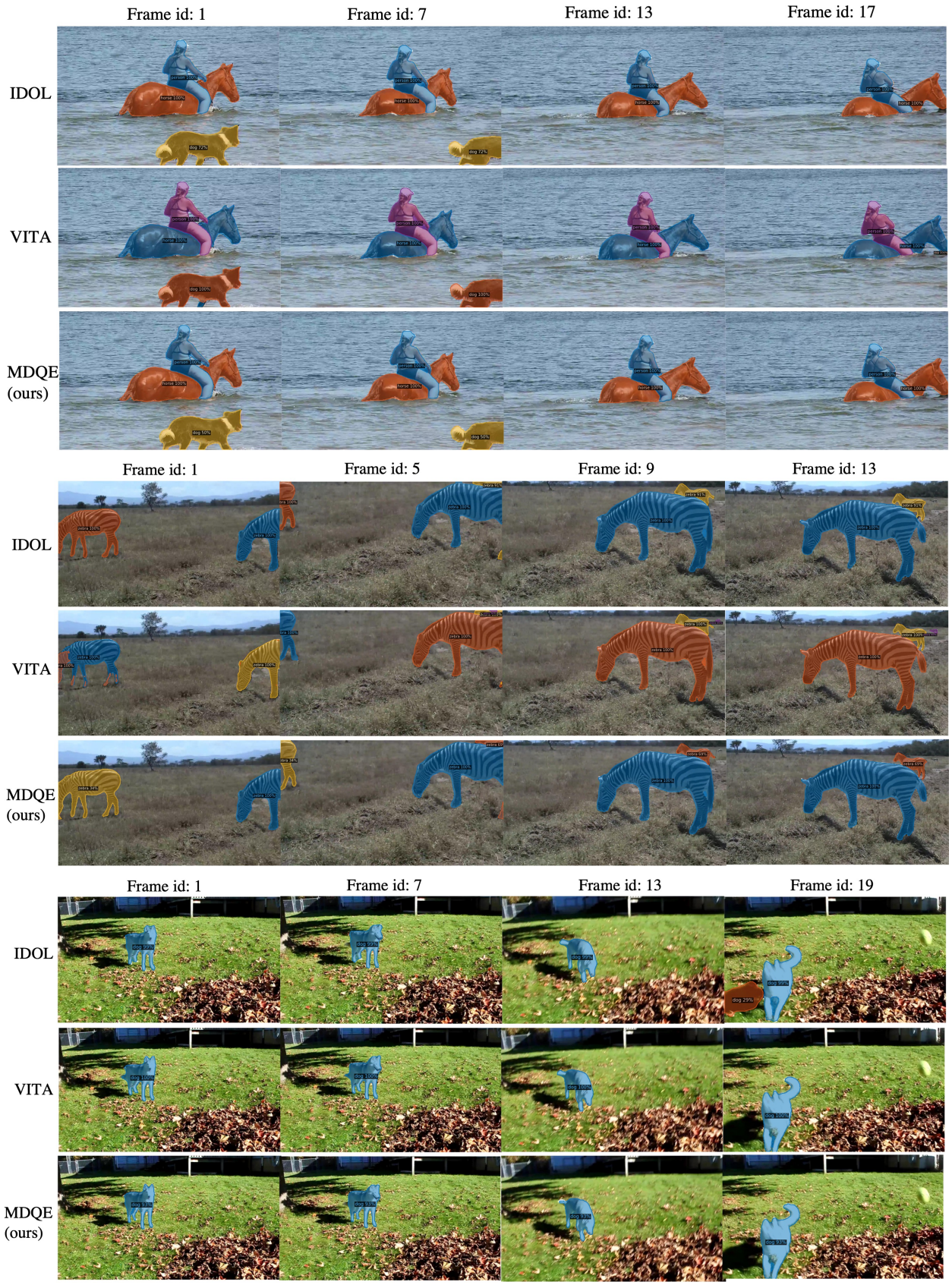

Figure 2. Visualization of instance masks on videos from the YouTube-VIS 2021 valid set.

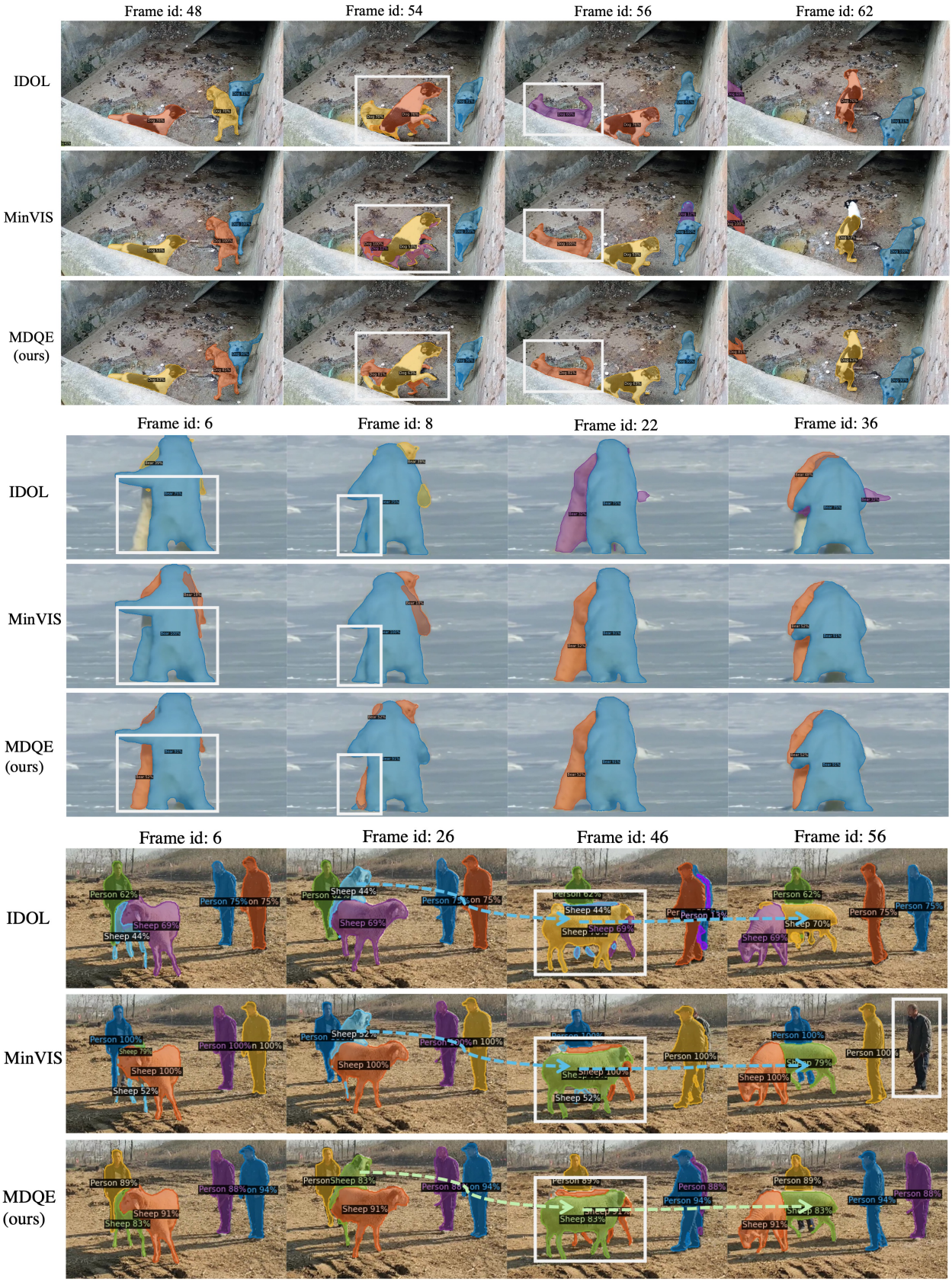

Figure 3. Visualization of instance masks on the OVIS valid set.

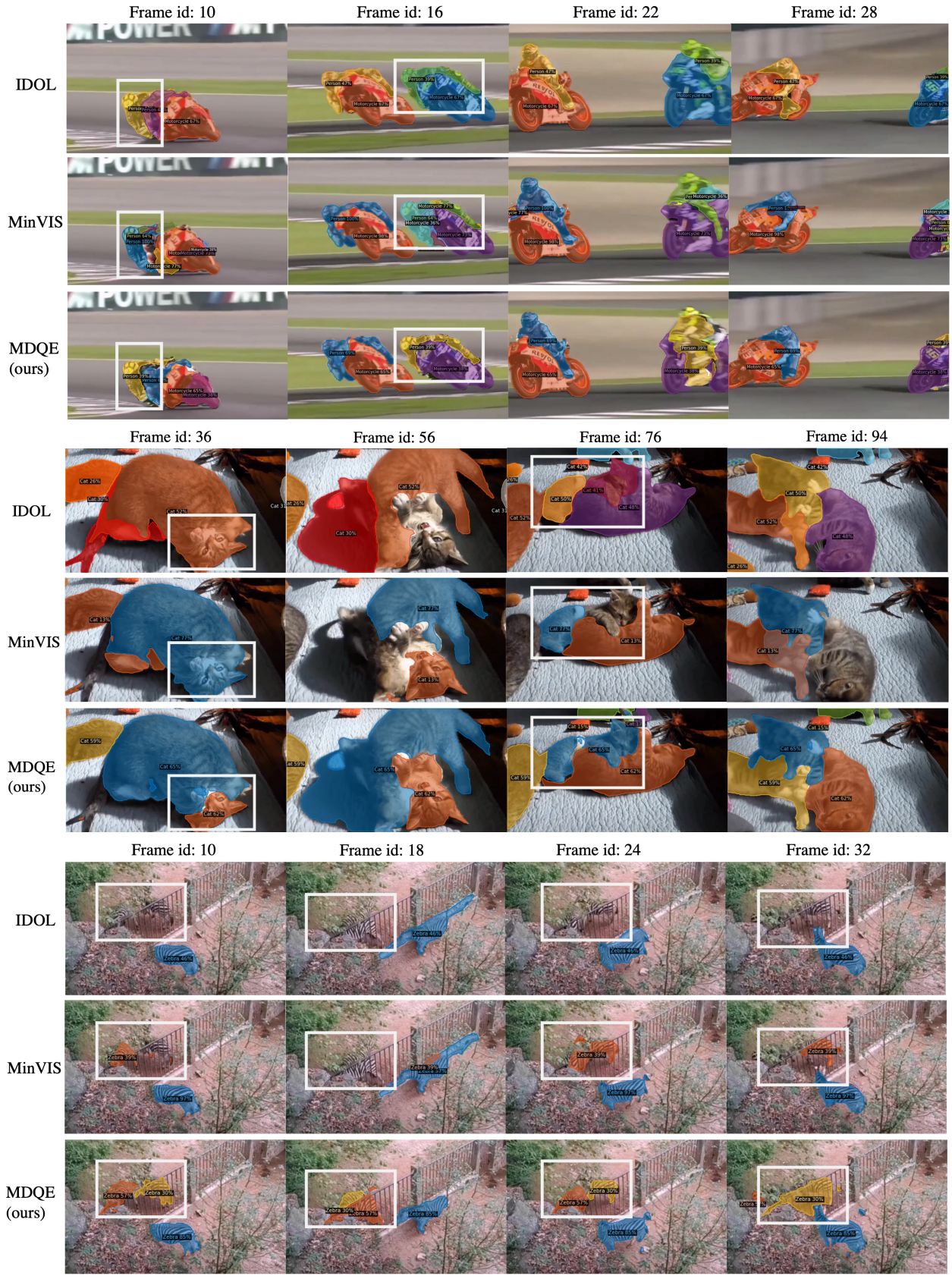

Figure 4. Visualization of instance masks on the OVIS valid set.

References

[1] Ali Athar, Sabarinath Mahadevan, Aljoša Ošep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *Eur. Conf. Comput. Vis.*, 2020. 2

[2] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9739–9748, 2020. 2

[3] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. SipMask: Spatial information preservation for fast image and video instance segmentation. *arXiv preprint arXiv:2007.14772*, 2020. 2

[4] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *arXiv preprint arXiv:2206.04403*, 2022. 1, 2

[5] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Adv. Neural Inform. Process. Syst.*, 2022. 1, 2

[6] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *Adv. Neural Inform. Process. Syst.*, 34:13352–13363, 2021. 2

[7] Minghan Li, Shuai Li, Lida Li, and Lei Zhang. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11215–11224, 2021. 2

[8] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. *arXiv preprint arXiv:2103.13746*, 2021. 2

[9] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9816–9825, 2021. 2

[10] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[11] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *Eur. Conf. Comput. Vis.*, 2022. 1, 2

[12] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *Eur. Conf. Comput. Vis.*, 2022. 1, 2

[13] Jialian Wu, Sudhir Yarram, Hui Liang, Tian Lan, Junsong Yuan, Jayan Eledath, and Gerard Medioni. Efficient video instance segmentation via tracklet query and proposal. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 959–968, 2022. 2

[14] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Int. Conf. Comput. Vis.*, pages 5188–5197, 2019. 2

[15] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. In *Int. Conf. Comput. Vis.*, pages 8043–8052, 2021. 2