CAVIS: Context-Aware Video Instance Segmentation (Supplementary Materials)

Seunghun Lee*, Jiwan Seo*, Kiljoon Han, Minwoo Choi, and Sunghoon Im
DGIST, Daegu, Korea
{lsh5688, eccaron, kiljoon_h, subminu, sunghoonim}@dgist.ac.kr

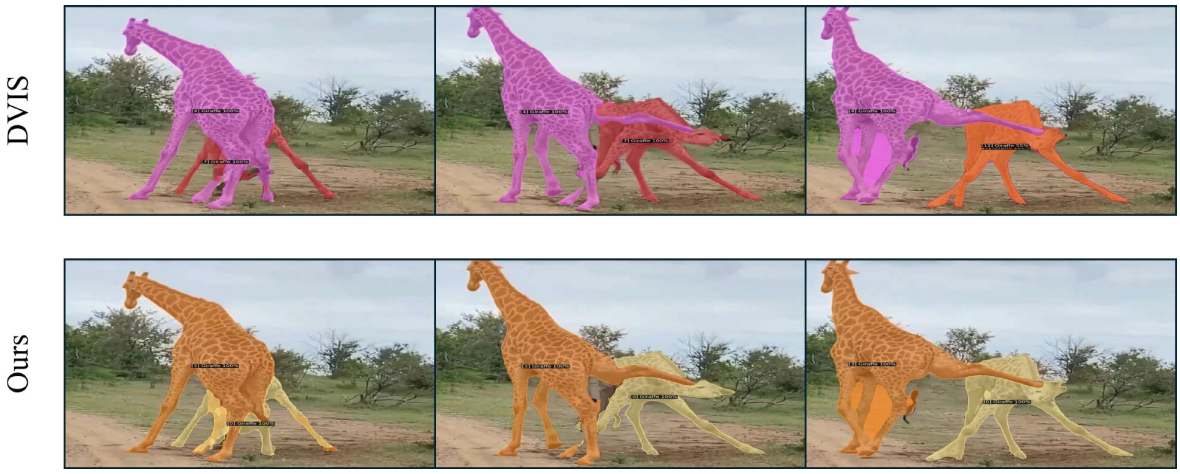

Figure 1. Potential error due to inaccurate mask predictions from the segmentation network.

1. Limitation

Video Instance Segmentation (VIS) is an advanced technology designed to perform segmentation and tracking concurrently, capturing the trajectories of individual instances within a video. While this technology has significant benefits, it also poses potential risks if misused, particularly in surveillance applications. Such misuse could lead to severe privacy infringements. It is important to note, however, that the dataset used in this study is a standard one within the VIS community and does not include any sensitive or personal information. This precaution helps mitigate the risk of our trained model being used for harmful purposes. Nonetheless, the potential for negative impacts should not be underestimated, and ethical considerations must guide the deployment of VIS technologies.

Potential error in prediction. Our model is designed to improve tracking accuracy by achieving precise object matching across frames rather than focusing on segmentation performance. Consequently, if the pretrained segmentation network produces inaccurate segmentation results, performance may decrease. However, even in scenarios with imprecise mask predictions, our proposed context-aware modeling can robustly track objects, as demonstrated in Fig. 1.

2. Experimental Details

2.1. Datasets

Youtube-VIS 2019 and 2021 YouTube-VIS was introduced by Yang et al. in their pioneering study on the VIS task [12]. This dataset comprises high-resolution YouTube videos, categorized into 40 distinct classes. The 2019 version of the dataset includes 2,238 videos for training, 302 for validation, and 343 for testing [12]. The 2021 update expands these numbers to 2,985, 421, and 453 videos for training, validation, and testing, respectively [13]. YouTube-VIS is utilized across various pixel-level video understanding tasks, including VIS, video semantic segmentation, and video object detection.

*These authors contribute equally to this work.

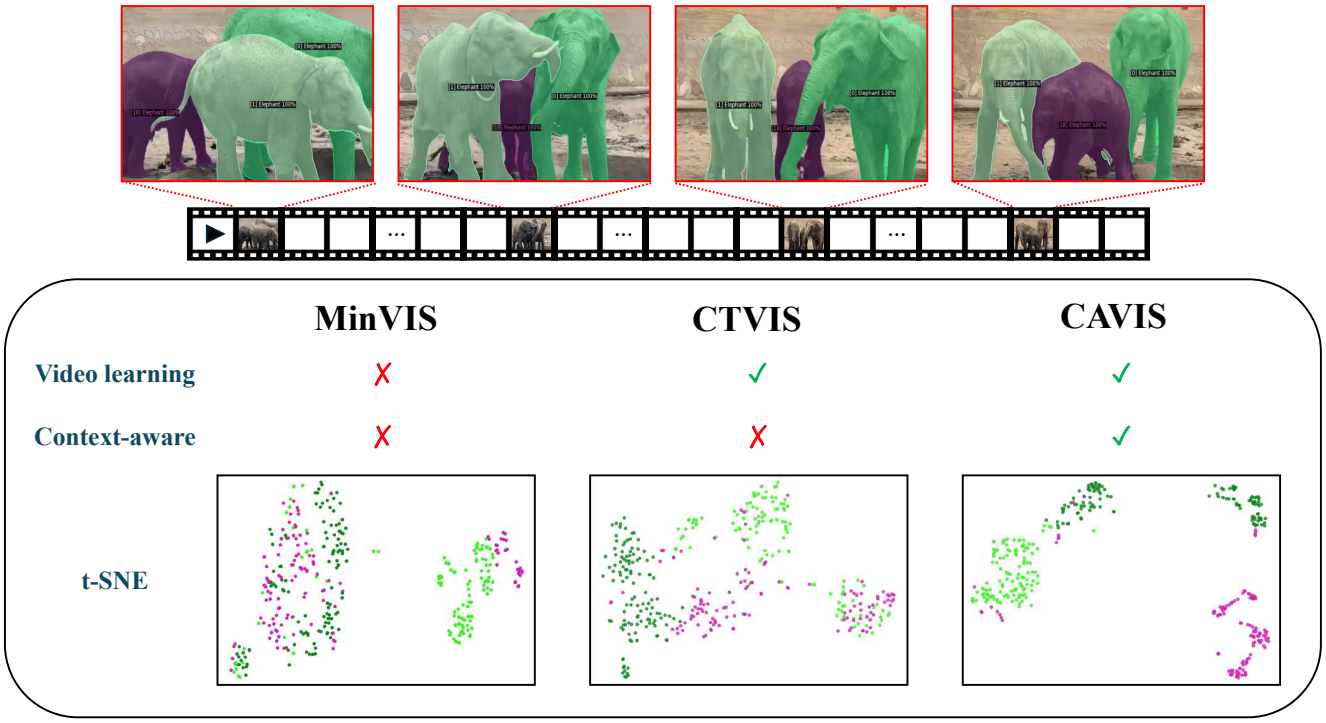

Figure 2. Visualization of object embeddings. Each point on the t-SNE [10] plot represents the learned object embeddings. The three different colors of points indicate the embeddings of three different elephants throughout the entire video.

Figure 3. Comparison of VIS results for the video in Fig. 2. These results show that our model robustly tracks objects even in scenes with severe occlusions.

OVIS The OVIS dataset [9] presents a significant challenge with its frequent occlusions and a realistic representation of common everyday objects. This makes it highly relevant for real-world applications. OVIS videos are longer and contain more objects compared to those in YouTube-VIS, which increases the complexity of segmentation and tracking tasks. The dataset is organized into training, validation, and test sets, with 607, 140, and 154 videos, respectively.

VIPSeg VIPSeg [8] is a comprehensive Video Panoptic Segmentation dataset that includes 3,536 videos and 84,750 frames, annotated with pixel-level panoptic labels. Unlike earlier VPS datasets that primarily focus on street views, VIPSeg offers a broader range of challenges and practical scenarios. It features 232 diverse settings and is annotated with 58 'thing' classes and 66 'stuff' classes, making it one of the most diverse and challenging datasets available in the field.

2.2. Implementation

Our segmentation approach employs the Mask2Former architecture [1], utilizing the officially recommended hyperparameters. For all experimental settings, we follow established practices by incorporating COCO joint training, as adopted in previous methodologies [3, 4, 11, 14, 15]. The tracking network consists of six transformer blocks. Within the tracking network's transformer blocks, we innovate by replacing the standard cross-attention layer with the referring cross-attention layer, as introduced in [15]. Additionally,

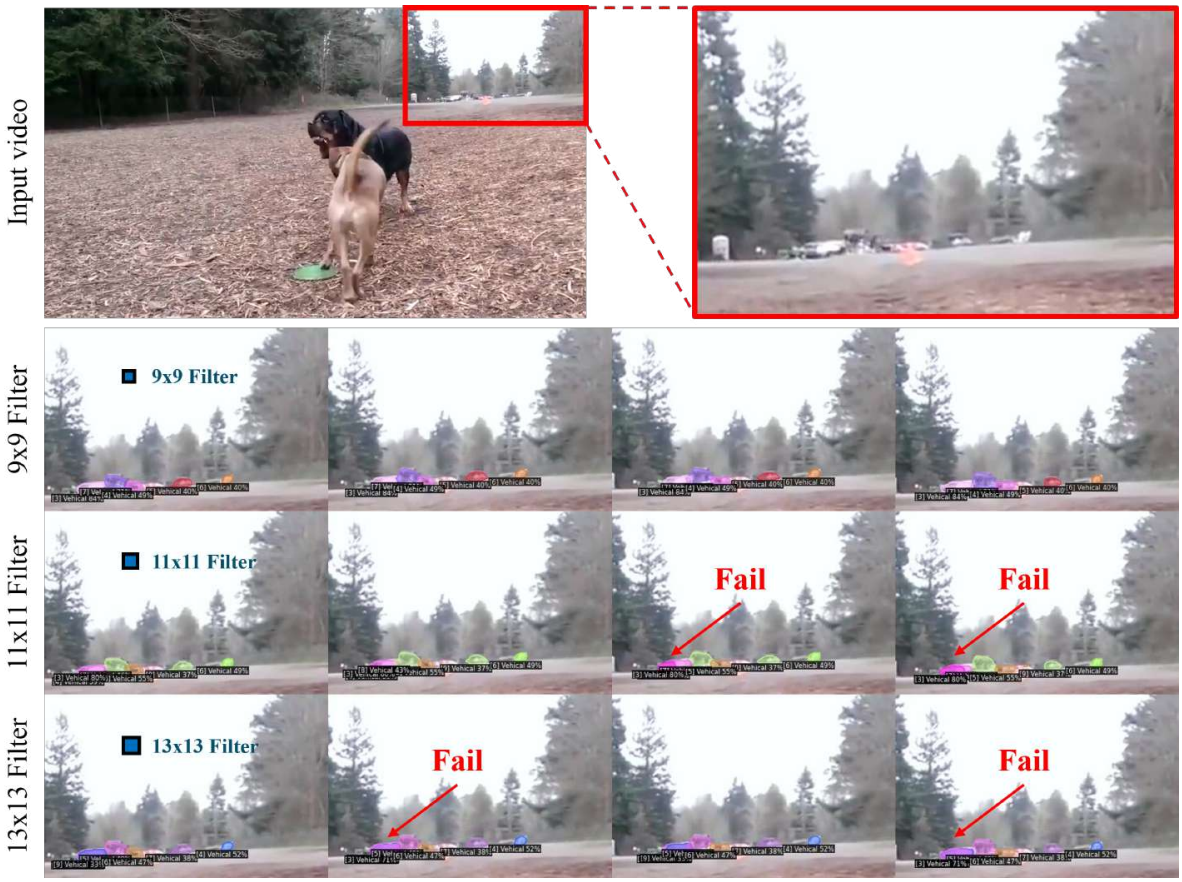

Figure 4. VIS results with various filter sizes.

we conduct experiments with the temporal refiner [15] over 160k iterations, specifically analyzing sequences of 15 consecutive frames to enhance tracking accuracy.

For efficient training, we adopt a staged approach where the segmentation network is trained first, followed by the tracking network with all other parameters frozen, promoting stability and efficiency in learning, as suggested by previous studies [6, 15]. Optimization is carried out using the AdamW optimizer [7], with a starting learning rate of 1e-4 and a weight decay of 5e-2. The training process spans 40k iterations for the segmentation network and 160k iterations for the tracking network, with learning rate reductions scheduled at 28k and 112k iterations, respectively. During training, we sample three frames for the segmentation network and five frames for the tracking network from each of eight batched videos. These frames undergo resizing to ensure the shorter side is between 320 and 640 pixels, while the longer side does not exceed 768 pixels. The loss function weights are set to $\lambda_{\text{cls}} = 2.0$, $\lambda_{\text{bce}} = 5.0$, $\lambda_{\text{dice}} = 5.0$, $\lambda_{\text{ctx}} = 2.0$, and $\lambda_{\text{pro}} = 2.0$ to balance the contributions of each component during training. For inference, the shorter side of input frames is scaled down to 448 pixels to maintain a consistent aspect ratio across inputs. All experiments are conducted using 8 RTX2080Ti GPUs for the ResNet-50 backbone and 8 RTX3090 GPUs for the Swin-L and ViT-L backbones, ensuring adequate computational resources are available for the demands of each model configuration.

3. Further Studies

Analysis on object embeddings. To demonstrate the effectiveness of our context-aware instance learning, we compare the distribution of object embeddings from three different models, as shown in Fig. 2. MinVIS does not engage in video learning, resulting in less effective distinction between objects. Compared to MinVIS, CTVIS shows a clearer object distinction by employing contrastive learning among object embeddings, but it still exhibits some overlaps in object clusters. In contrast, CAVIS forms much more distinct object clusters, highlighting the advantage of leveraging contextual information for object identification. This trends are reflected in the VIS results, as shown in Fig. 3.

Effective filter size. Videos often contain objects of varying sizes, and for smaller objects, using an excessively large context area can introduce noise, leading to inaccurate match-

Table 1. Ablation studies on each component of CAVIS. (a-d) present the results from the segmentation network, while the others present those from the tracking network. "CL" denotes contrastive learning.

(a) Context-aware feature learning, PCC loss

	CL with $\hat{Q}$	$\mathcal{L}_{\text{CTX}}$	$\mathcal{L}_{\text{PCC}}$	AP
(i)				26.4
(ii)	✓			27.9
(iii)		✓		29.1
(iv)			✓	27.6
(v)	✓		✓	28.3
(vi)		✓	✓	**29.5**

(b) Context filter size

Filter size	AP
3	27.3
5	28.3
7	28.7
9	**29.5**
11	28.9

(c) Sampled frames

# of frames	AP
2	29.5
3	**30.0**
4	28.7

(d) Context filter type

Metric	Context filter type	
	Average	Learnable
AP	**29.5**	28.4

(e) Cross-Attention for $\mathcal{T}$

Metric	Cross-Attention	
	$\hat{Q}$	Q
AP	34.4	**36.1**

(f) Context alignment

Metric	Context alignment	
	✗	✓
AP	32.8	**36.1**

Figure 5. VIS results from our model on a video containing a fast-moving object.

ing as shown in Fig. 4. To better understand this effect, we analyze the impact of different filter sizes to identify the optimal value. Our findings indicate that the overall trend remains consistent, regardless of variations in the number of frames used during training.

Ablation study with minimal setups. To simplify reproducibility, we additionally provide ablation studies on the OVIS dataset [9] with the ResNet-50 [2] backbone, detailed in Tab. 1. The results exhibit similar trends to those observed in the main text, further validating the consistency of our findings. For these experiments, we train the segmentation network with 2 frames over 40k iterations, while the tracking network is trained with 5 frames over 40k iterations. Experiments (i- iii) show that implementing contrastive learning, whether with standard or context-aware instance features, leads to significant performance gains. Particularly, context-aware instance features result in a notable +2.7 AP improvement over the baseline, a considerable increase compared to the +1.3 AP improvement observed with standard instance features.

Robustness of our model. Our method does not rely solely on context. By incorporating both context and instance features, our approach shows robustness even in scenes containing fast-moving objects where context changes rapidly, as shown in Fig. 5.

Performance on long video. We additionally report the performance on the YouTube-VIS 2022 dataset, a well-known benchmark featuring long video sequences. Its validation set includes 71 additional videos compared to the YouTube-VIS 2021 dataset, making it particularly challenging due to the need for accurately tracking dynamically appearing and disappearing objects over extended periods. We evaluate our model on these 71 long videos and compare it against existing state-of-the-art models with a ResNet-50 backbone. As shown in Tab. 2, our approach outperforms existing methods, demonstrating that our context-aware modeling remains effective for robust object matching even in long-range video scenarios.

Method	AP
MinVIS [5]	23.3
DVIS [15]	31.6
VITA [3]	32.6
DVIS++ [16]	37.2
GenVIS [4]	37.5
Ours	**38.6**

Table 2. Comparison on YTVIS 2022 dataset.

Method	Time (ms)	YTVIS19	OVIS
DVIS [15]	**78.9**	51.2	30.2
GenVIS [4]	80.1	50.0	35.8
Ours	85.6	**55.7**	**37.6**

Table 3. Computational cost.

Computational cost. We compare the inference speed of our approach against recent state-of-the-art methods, DVIS

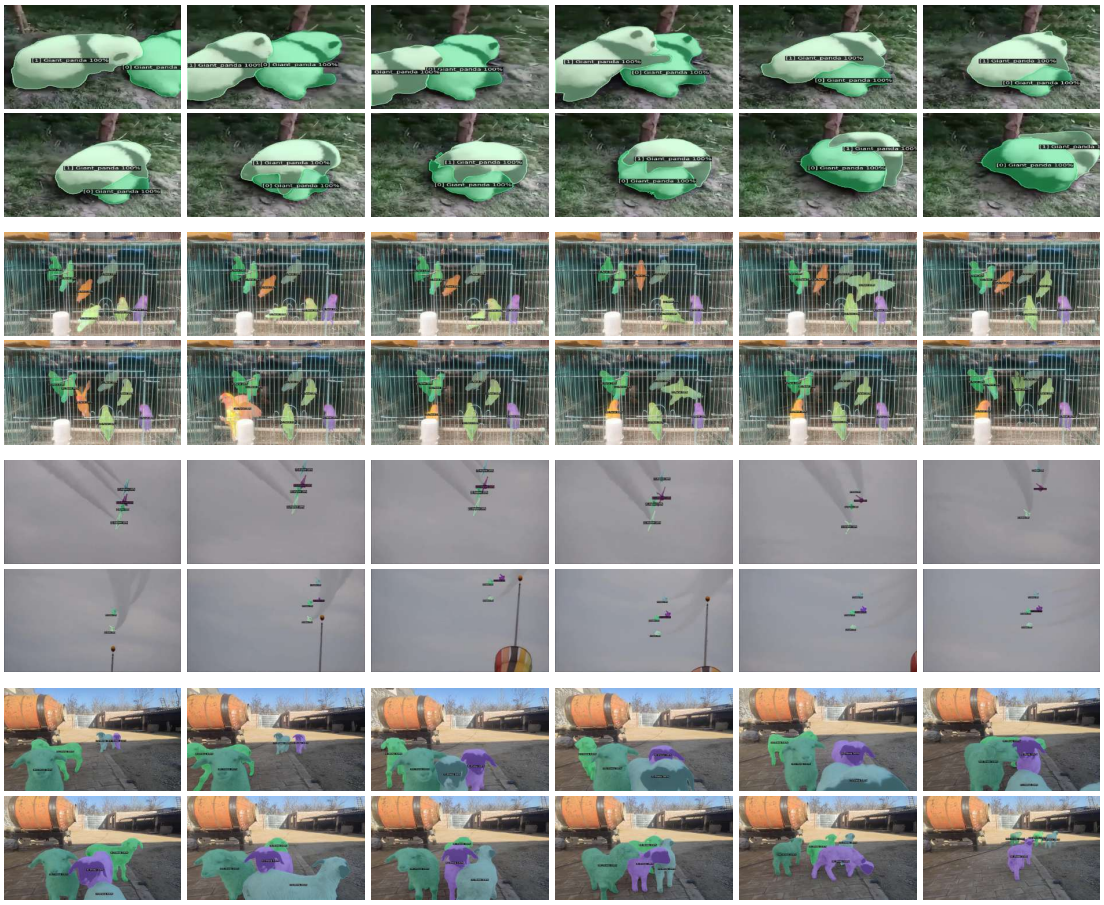

Figure 6. Additional qualitative results on OVIS dataset.

[15] and GenVIS [4], as shown in Tab. 3. The inference speeds were measured under identical conditions on a 2080ti GPU using the ResNet-50 backbone. Our method requires an additional time cost of 6.7ms and 5.5ms compared to DVIS and GenVIS, respectively. However, this cost is justified by performance gains of +1.8 AP and +7.6 AP on OVIS, and +4.5 AP and +5.7 AP on YTVIS19, demonstrating a reasonable trade-off between increased computation and improved accuracy.

Additional qualitative results. We provide additional qualitative results of CAVIS across various datasets, as depicted in Fig. 6-9. These results underscore the robust capability of CAVIS to track objects in diverse scenarios for both VIS and VPS tasks. Notably, CAVIS excels in environments featuring numerous similar objects, fast-moving objects, and significant occlusions, demonstrating its effectiveness across complex dynamic scenes.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[3] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *Advances in Neural Information Processing Systems*, 35:23109–23120, 2022. 2, 4

[4] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14623–14632, 2023. 2, 4, 5

[5] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Advances in Neural Information Processing Systems*, 35:31265–31277, 2022. 4

[6] Junlong Li, Bingyao Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. Tcovis: Temporally consistent online video instance

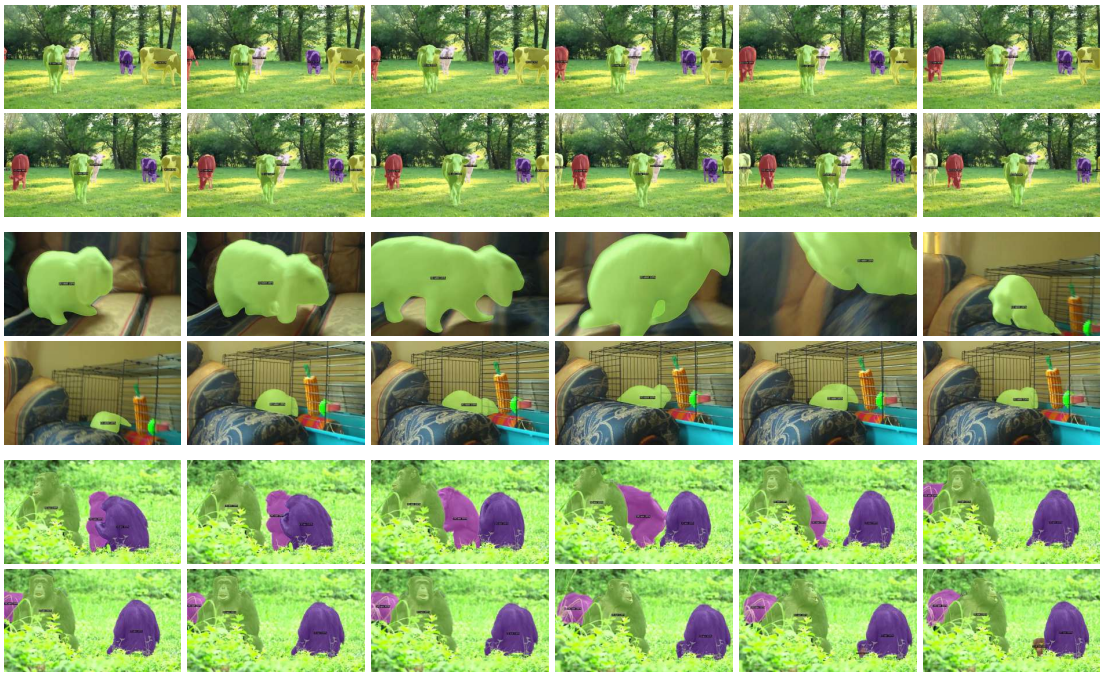

Figure 7. Additional qualitative results on Youtube-VIS 2019 dataset.

Figure 8. Additional qualitative results on Youtube-VIS 2021 dataset.

segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1097–1107, 2023. 3

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[8] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21033–21043, 2022. 2

[9] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu,

Figure 9. Additional qualitative results on VIPSeg dataset.

Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039, 2022. 2, 4

[10] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2

[11] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *European Conference on Computer Vision*, pages 553–569. Springer, 2022. 2

[12] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 1

[13] Linjie Yang, Yuchen Fan, Yang Fu, and Ning Xu. The 3rd large-scale video object segmentation challenge - video instance segmentation track, 2021. 1

[14] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. Ctvis: Consistent training for online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 899–908, 2023. 2

[15] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. *arXiv preprint arXiv:2306.03413*, 2023. 2, 3, 4, 5

[16] Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *arXiv preprint arXiv:2312.13305*, 2023. 4