AutoQ-VIS: Improving Unsupervised Video Instance Segmentation via Automatic Quality Assessment

Kaixuan Lu[1] Mehmet Onurcan Kaya[1,2] Dim P. Papadopoulos[1,2]
[1]Technical University of Denmark [2]Pioneer Centre for AI
s232248@student.dtu.dk, monka@dtu.dk, dimp@dtu.dk

Abstract

Video Instance Segmentation (VIS) faces significant annotation challenges due to its dual requirements of pixel-level masks and temporal consistency labels. While recent unsupervised methods like VideoCutLER eliminate optical flow dependencies through synthetic data, they remain constrained by the synthetic-to-real domain gap. We present AutoQ-VIS, a novel unsupervised framework that bridges this gap through quality-guided self-training. Our approach establishes a closed-loop system between pseudo-label generation and automatic quality assessment, enabling progressive adaptation from synthetic to real videos. Experiments demonstrate state-of-the-art performance with 52.6 AP_{50} on YouTubeVIS-2019 `val` set, surpassing the previous state-of-the-art VideoCutLER by 4.4%, while requiring no human annotations. This demonstrates the viability of quality-aware self-training for unsupervised VIS. The source code of our method is available at here.

1. Introduction

Video Instance Segmentation (VIS) is the challenging task of simultaneously detecting, segmenting, and tracking object instances across video sequences [8, 11, 16, 17, 20, 23]. This capability is fundamental for scene understanding in applications ranging from autonomous driving [21] to video content editing [24]. However, training high-performance VIS models typically requires pixel-level annotations across all frames [20]. This process is expensive due to the labor-intensive nature of annotating temporal consistency and instance identities. As a result, there is an urgent need to develop unsupervised video instance segmentation frameworks that can accurately interpret video content and function effectively across diverse, unlabeled environments.

Prior work [1, 4, 6, 10, 12, 19, 22] in unsupervised video segmentation predominantly addresses Video Object Segmentation (VOS), focusing on separating a single foreground object via motion or consistency cues. While

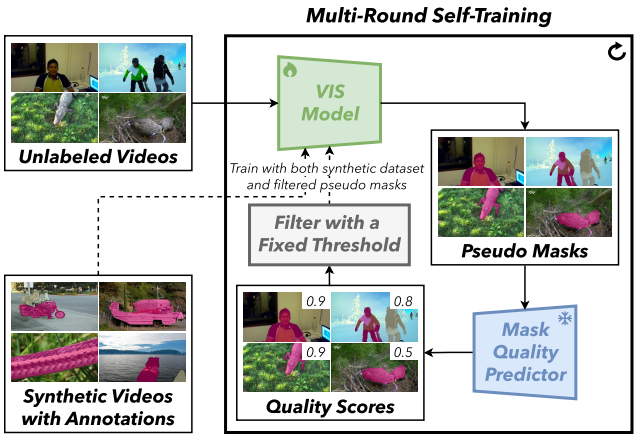

Figure 1. **AutoQ-VIS overview.** In the initial training stage, both the VIS model and the mask quality predictor are trained on synthetic videos with pseudo annotations [15]. During the multi-round self-training stage, the VIS model generates pseudo masks on unlabeled videos, which are then scored by the frozen quality predictor. Pseudo masks with high predicted quality are selected and added to the training set. The VIS model is subsequently re-trained on both the synthetic data and the selected pseudo labels, enabling iterative refinement and progressive performance gains.

OCLR [18] introduces unsupervised VIS that supports multiple instances, its predefined object count during training prevents dynamic adaptation to varying instances during inference. Furthermore, prior approaches [4, 10, 12, 18, 19] rely on optical flow estimators (e.g., RAFT [13]) that are trained on human-annotated datasets. VideoCutLER [15] marks a breakthrough in unsupervised VIS and achieves unprecedented performance by demonstrating multi-instance segmentation without optical flow dependencies. Its core innovation lies in synthetic video generation via spatial augmentations of CutLER [14] pseudo-labels from ImageNet [5]. However, VideoCutLER remains constrained by synthetic-to-real domain gaps and static instance modeling, i.e., the synthetic videos lack natural and realistic motion patterns.

Building upon VideoCutLER's synthetic data paradigm, which generates training videos through spatial augmentations of static image pseudo-labels, we introduce AutoQ-VIS to address its critical domain gap limitation. While VideoCutLER's synthetic videos provide initial instance awareness, they lack natural motion patterns and real-world appearance variations, hindering adaptation to authentic video dynamics. Our framework bridges this synthetic-to-real domain gap through a self-training loop that progressively adds quality-filtered pseudo-labels from unlabeled real videos. Inspired by Mask Scoring R-CNN [9], which directly predicts mask quality scores via an auxiliary branch, we implement a quality assessment module for pseudo-label filtering of instance masks.

AutoQ-VIS advances unsupervised video instance segmentation through an iterative self-training paradigm with quality-aware pseudo-label selection (Fig. 1). The system initializes using synthetic video data from VideoCutLER, which provides pseudo-labels to bootstrap a VideoMask2Former [2, 3] VIS model and a specialized mask quality predictor (Sec. 2.1). The mask quality predictor estimates mask IoU quality scores by analyzing frame-level features and mask predictions. During multi-round optimization, the VIS model generates pseudo-labels on unlabeled videos, which are then scored by the quality predictor. High-quality pseudo-labels surpassing a fixed threshold are progressively incorporated into the training set, enabling dataset augmentation without any human supervision (Sec. 2.2). To enhance the mask head training, we employ a DropLoss that zeros out mask losses whose maximum ground-truth overlap falls below 0.01 (Sec. 2.3). By alternating rounds of VIS training (with occasional weight resets) and quality-based dataset expansion, AutoQ-VIS dynamically enriches its training dataset and steadily sharpens segmentation performance.

Our key contributions are threefold: **(1) Annotation-Free VIS Framework:** We propose AutoQ-VIS, an unsupervised framework that overcomes annotation dependency through cyclic pseudo-label refinement with automated quality control, enabling video instance segmentation training directly from unlabeled videos. **(2) Automatic Quality Assessment:** We propose a simple quality predictor that reliably filters pseudo labels across self-training rounds. **(3) New State-of-the-art Performance:** Our AutoQ-VIS archives 52.6 AP_{50} on YouTubeVIS-2019 [20] `val` split, surpassing the previous state-of-the-art VideoCutLER [15] by 4.4 AP_{50}.

2. Method

AutoQ-VIS operates through three stages: (1) **Initial Training** (Sec. 2.1): Jointly pretrain VideoMask2Former and the mask quality predictor on VideoCutLER's synthetic videos; (2) **Multi-Round Self-Training** (Sec. 2.2): Iteratively generate pseudo-labels on real unlabeled videos, filter via quality scores, and augment training data; (3) **DropLoss** (Sec. 2.3): Suppress low-IoU mask predictions to enhance mask head training. This quality-guided pipeline progressively improves segmentation accuracy without any human annotations.

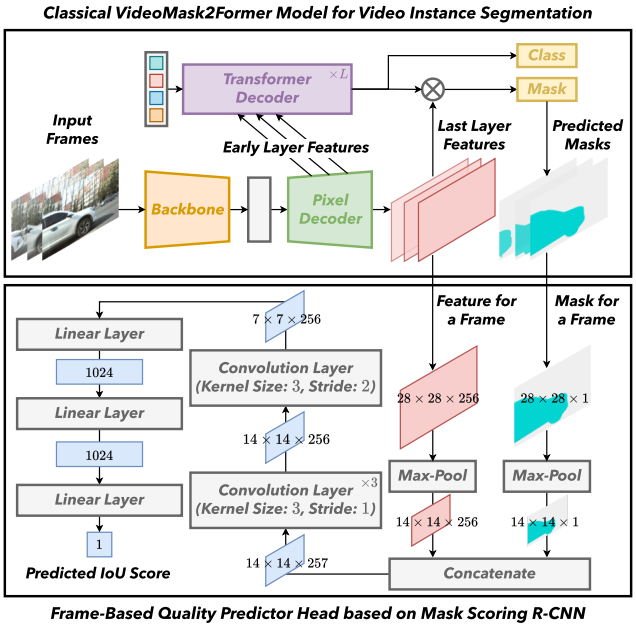

Figure 2. **Network architecture of VideoMask2Former [2, 3] and Mask Quality Predictor.** Our quality predictor integrates mask predictions and pixel decoder features following [9], employing a sequential architecture with four convolution layers (3×3 kernels, final layer stride of 2 for spatial reduction) followed by three fully-connected layers that ultimately produce mask IoU predictions.

2.1. Initial training stage

Video instance segmentation (VIS) model. Following VideoCutLER [15], we use the VideoMask2Former [2, 3] with the ResNet-50 [7] backbone as our video instance segmentation (VIS) model.

Quality predictor. For the quality predictor, as shown in Fig. 2, we use an architecture inspired by Mask Scoring R-CNN [9]. Our architecture processes individual frame features and single-object mask predictions per inference step. Supervision is established through threshold-binarized (0.5) mask IoU between predictions and matched ground truths, optimized via ℓ_2 regression loss.

Synthetic videos. VideoCutLER [15] provides a high-quality pseudo-labeled synthetic video dataset that was built on the unlabeled images from ImageNet [5], which is very suitable to train and initialize our VIS model and quality predictor. We also use the trained VideoMask2Former model [2] from VideoCutLER to initialize our VIS model.

Method	AP_{50}	AP_{75}	AP	AP_S	AP_M	AP_L	AR_{10}
MotionGroup [19]	0.5	0.0	0.1	0.0	0.4	0.1	1.2
OCLR [18]	5.5	0.3	1.6	0.1	1.6	6.1	11.5
CutLER [14]	36.4	13.5	16.0	3.5	13.9	26.0	29.8
VideoCutLER [15]	48.2	22.9	24.5	6.7	17.7	36.3	42.3
AutoQ-VIS	52.6	28.2	28.1	6.7	21.2	40.7	42.5
vs. previous SOTA	+4.4	+5.3	+3.6	+0.0	+3.5	+4.4	+0.2

Table 1. **YouTubeVIS-2019 `val`.** We reproduced MotionGroup [19], OCLR [18], CutLER [14], and VideoCutLER [15] results with the official code and checkpoints. AutoQ-VIS outperforms the state-of-the-art VideoCutLER by 4.4 AP_{50}. We evaluate results on YouTubeVIS-2019's `val` split in a class-agnostic manner.

2.2. Multi-round self-training

As shown in Fig. 1, we optimize the VIS model through iterative self-training and dynamic dataset augmentation. The training dataset is initialized using the synthetic videos from VideoCutLER [15]. Empirically, we find that executing model parameter restoration from the initial model weight (trained in Sec. 2.1) achieves a better performance.

After training the VIS model, we use the predicted masks on unlabeled videos with a confidence score over 0.25 as pseudo-labels. Then we use the quality predictor to predict the IoU of each pseudo-label predicted mask. Let $\hat{\text{IoU}}_l$ denote the predicted IoU of label l, s_l denotes the confidence score of label l. We define the quality score of label l as $Q_l = \hat{\text{IoU}}_l \cdot s_l$.

We implement quality-based pseudo-label selection using a fixed quality score threshold τ_{th}. For each pseudo-label l, we select it if $Q_l \geq \tau_{th}$. In the end of each round, we add all the pseudo-labels to the training dataset.

2.3. DropLoss for mask head

We enhance the mask head training by suppressing loss contributions from low-overlap predictions, following CutLER [14]. For each predicted mask m_i, we discard its loss contribution if its maximum ground truth IoU falls below the threshold τ^{IoU}:

$$\mathcal{L}_{\text{drop}}(m_i) = \mathbb{1}(\text{IoU}_i^{\text{max}} > \tau^{\text{IoU}})\mathcal{L}_{\text{vanilla}}(m_i) \quad (1)$$

Here, $\text{IoU}_i^{\text{max}}$ is the maximum overlap between m_i and any ground truth mask, while $\mathcal{L}_{\text{vanilla}}$ denotes the original mask head loss from VideoMask2Former [2, 3]. We employ a low threshold ($\tau^{\text{IoU}} = 0.01$) to filter only near-zero overlap predictions.

3. Experiments

Datasets. Our model is trained on synthetic videos from VideoCutLER [15] and the unlabeled `train` split of YouTubeVIS-2019 [20]. We evaluate our model's performance on the `val` split of YouTubeVIS-2019 in a class-agnostic manner.

Method	AP_{50}	AP_{75}	AP	AP_S	AP_M	AP_L	AR_{10}
Theoretical limit	76.8	48.7	46.8	13.5	43.6	62.9	58.0
Practical limit	62.7	33.2	33.9	4.3	27.3	53.2	47.5
AutoQ-VIS	52.6	28.2	28.1	6.7	21.2	40.7	42.5

Table 2. **Comparison with the theoretical and practical limit.** *Theoretical Limit*: Upper-bound performance achieved by training on ground-truth labels from YouTubeVIS-2019 `train` split in class-agnostic mode, representing ideal supervision conditions. *Practical Limit*: Best attainable performance when using all pseudo-labels with IoU $\geq$ 0.5 against ground truth, simulating perfect pseudo-label selection.

Evaluation metrics. Following [15], we use Average Precision (AP) and Average Recall (AR) as evaluation metrics. We evaluate the models in a class-agnostic manner, treating all classes as a single one during evaluation.

Implementation details. For the initial training stage, we use the pretrained VideoMask2Former model [2, 3] from VideoCutLER [15] to initialize our VIS model. Then we train the VIS model and quality predictor on synthetic videos from VideoCutLER for 8,000 iterations using a single V100 GPU, with a batch size of 2 and a learning rate of 2×10^{-5}. For each round of multi-round self-training, we train the VIS model for 10,000 iterations on two V100 GPUs, with a batch size of 4 and a learning rate of 2×10^{-5}. In practice, we find that two rounds of self-training provide the best performance.

3.1. Experimental results

Comparison with the-state-of-the-art method. We compare AutoQ-VIS with previous top-performing methods in Tab. 1. AutoQ-VIS achieves a remarkable improvement (about 4.4% AP_{50}). Especially for AP_{75}, AutoQ-VIS can outperform the state-of-the-art VideoCutLER [15] by 5.3%.

Comparison with the theoretical and practical limit. Tab. 2 reveals a significant performance gap (10.1 AP_{50}) between AutoQ-VIS and the practical upper bound, indicating substantial potential for improvement through enhanced pseudo-label utilization. The theoretical limit represents a fully supervised training using all ground-truth annotations from YouTubeVIS-2019 `train` set. The practical limit represents an oracle experiment that simulates perfect pseudo-label selection by using all predictions with IoU $\geq$ 0.5 against ground truth.

Qualitative results. Fig. 3 demonstrates AutoQ-VIS's advancements over VideoCutLER [15]. As we observe, AutoQ-VIS is capable of discovering more objects and producing higher-quality segmentation masks.

3.2. Ablation studies

Component ablation analysis. Tab. 3 quantifies individual component contributions through progressive additions. The DropLoss provides the most substantial gains (+4.6

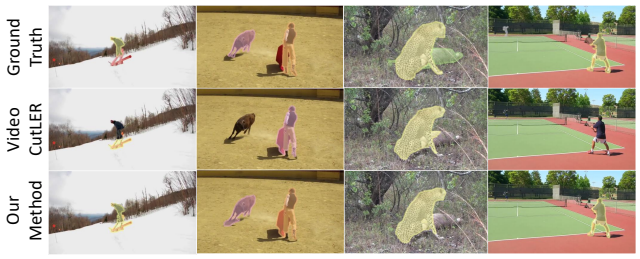

Figure 3. **The qualitative results on YouTubeVIS-2019 `val` split.** AutoQ-VIS demonstrates superior instance discovery capabilities compared to VideoCutLER [15]: (1) Enhanced multi-object detection capacity, particularly for semantically distinct instances (e.g., person and bull in Column 2); (2) Improved segmentation fidelity through precise boundary delineation (e.g., the leopard in Column 3). (3) Better comprehensive instance coverage, eliminating false negatives (e.g., detecting humans in Columns 1 & 4 that VideoCutLER completely misses, even without occlusion or scale challenges).

Method	AP_{50}	AP_{75}	AP	AP_S	AP_M	AP_L	AR_{10}
w/o quality predictor	50.5	25.9	27.2	5.7	19.0	40.5	**43.3**
w/o DropLoss	48.0	23.7	24.6	3.8	16.9	37.8	39.2
w/o resetting each round	51.6	28.0	**28.2**	6.4	**22.8**	40.6	42.9
AutoQ-VIS	**52.6**	**28.2**	28.1	**6.7**	21.2	**40.7**	42.5

Table 3. **Ablation study on the contribution of each component.** *Without quality predictor:* We remove the quality predictor, and use the confidence score s_l as quality score Q_l with threshold $\tau_{th} = 0.85$. *Without DropLoss:* We use the vanilla loss for the mask head instead of the DropLoss. *Without resetting each round:* The model weights are not reset at the beginning of each round.

Self-training	AP_{50}	AP_{75}	AP	AP_S	AP_M	AP_L	AR_{10}
1 round	51.3	26.5	26.9	**7.0**	**22.3**	38.4	**43.2**
2 rounds	**52.6**	**28.2**	**28.1**	6.7	21.2	**40.7**	42.5
3 rounds	52.0	27.0	27.7	6.2	21.8	40.1	42.1

Table 4. **Ablation study on the number of self-training rounds.** Our framework achieves peak performance (52.6 AP_{50}) at the second round before gradual degradation (-0.6 AP_{50}) from pseudo-label noise accumulation. This establishes round 2 as the optimal stopping point to balance accuracy and error propagation risks.

AP_{50}), followed by the quality predictor's +2.1 AP_{50} improvement. Notably, even the confidence score baseline surpasses VideoCutLER by +2.3 AP_{50}, demonstrating fundamental advantages of our self-training method. While model resetting yields marginal gains (+1.0 AP_{50}, +0.2 AP_{75}), it maintains baseline AP performance.

Self-training round analysis. Tab. 4 tests how many self-training rounds work best. The model hits its peak (52.6 AP_{50}) at round 2, then slowly gets worse. This drop occurs because, as we perform more rounds, mistakes in the pseudo-labels pile up.

Quality score threshold τ_{th} analysis. Tab. 5 examines

τ_{th}	AP_{50}	AP_{75}	AP	AP_S	AP_M	AP_L	AR_{10}
0.95	48.7	25.3	26.1	5.9	18.8	38.5	40.7
0.85	48.8	24.6	26.0	5.8	18.6	38.6	41.2
0.75	**52.6**	**28.2**	**28.1**	**6.7**	**21.2**	**40.7**	**42.5**
0.50	52.4	25.7	27.1	6.5	20.1	39.9	42.2

Table 5. **Ablation study on the quality score threshold τ_{th}.** Optimal performance (52.6 AP_{50}) emerges at $\tau_{th} = 0.75$, balancing valid sample retention and noise suppression. Lower threshold ($\tau_{th} = 0.50$) degrades results by admitting too many low-quality predictions, while the higher thresholds ($\tau_{th} = 0.95$ and $\tau_{th} = 0.80$) oversuppress valid samples.

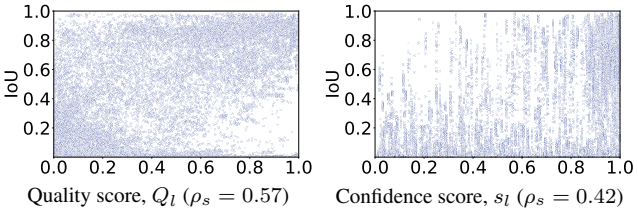

Figure 4. **Visualized comparison of quality score Q_l and confidence score s_l.** Here, ρ_s denotes the Spearman's rank correlation coefficient. Subplot (a) visualizes quality scores Q_l and their ground truth IoU. Subplot (b) visualizes confidence scores s_l and their ground truth IoU.

the impact of quality score thresholds on pseudo-label selection. We observe a non-monotonic relationship: While lower thresholds ($\tau_{th} \leq 0.75$) generally yield superior performance by retaining more valid samples, excessively lenient selection ($\tau_{th} = 0.50$) introduces noisy supervision, degrading results. The optimal balance occurs at $\tau_{th} = 0.75$, achieving peak performance.

Quality score vs. confidence score. As shown in Fig. 4, our quality score Q_s has a higher correlation to the IoU of the pseudo label and the ground truth label than the confidence score of VideoMask2Former, which proves our quality predictor's effectiveness in pseudo-label quality assessment. This results in a significant improvement in the final VIS model performance (+2.1 AP_{50}) in Tab. 3.

4. Conclusion

We present AutoQ-VIS, a quality-aware self-training framework that advances unsupervised video instance segmentation through iterative pseudo-label refinement with automatic quality control. By establishing a closed-loop system of pseudo-label generation and automatic quality assessment, our method achieves state-of-the-art performance (52.6 AP_{50}) on YouTubeVIS-2019 `val` split without requiring any human annotations. The simple quality predictor proves effective in pseudo-label quality assessment.

References

[1] Nikita Araslanov, Simone Schaub-Meyer, and Stefan Roth. Dense unsupervised learning for video segmentation. *Advances in Neural Information Processing Systems*, 34: 25308–25319, 2021. 1

[2] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 2, 3

[3] 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, 3

[4] Subhabrata Choudhury, Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Guess What Moves: Unsupervised Video and Image Segmentation by Anticipating Motion. In *British Machine Vision Conference (BMVC)*, 2022. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2

[6] Emanuela Haller and Marius Leordeanu. Unsupervised object segmentation in video by efficient selection of highly probable positive features. In *Proceedings of the IEEE international conference on computer vision*, pages 5085–5093, 2017. 1

[7] 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. 2

[8] 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. 1

[9] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6409–6418, 2019. 2

[10] Long Lian, Zhirong Wu, and Stella X Yu. Bootstrapping objectness from videos by relaxed common fate and visual grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14582–14591, 2023. 1

[11] Chung-Ching Lin, Ying Hung, Rogerio Feris, and Linglin He. Video instance segmentation tracking with a modified vae architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13147–13157, 2020. 1

[12] Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. Em-driven unsupervised learning for efficient motion segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4462–4473, 2022. 1

[13] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 1

[14] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3124–3134, 2023. 1, 3

[15] Xudong Wang, Ishan Misra, Ziyun Zeng, Rohit Girdhar, and Trevor Darrell. Videocutler: Surprisingly simple unsupervised video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22755–22764, 2024. 1, 2, 3, 4

[16] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8741–8750, 2021. 1

[17] 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. 1

[18] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. *Advances in neural information processing systems*, 35: 28023–28036, 2022. 1, 3

[19] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7177–7188, 2021. 1, 3

[20] 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, 2, 3

[21] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1

[22] Kaihua Zhang, Zicheng Zhao, Dong Liu, Qingshan Liu, and Bo Liu. Deep transport network for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8781–8790, 2021. 1

[23] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1282–1291, 2023. 1

[24] Tianfei Zhou, Fatih Porikli, David J Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7099–7122, 2022. 1