Introducing a self-supervised, superfeature-based network for video object segmentation

PhD student: Marcelo Mendonça
Postgraduate Program in Mechatronics
Federal University of Bahia
Salvador, Brazil
marceloms@ufba.br

Advisor: Luciano Oliveira
dept. of Computer Science
Federal University of Bahia
Salvador, Brazil
lrebouca@ufba.br

***Abstract*—This 3-page long paper summarizes our PhD thesis with the aim of participating in the CBIC23 Contest of Theses and Dissertations (CTD). Our thesis introduces a novel video-object segmentation (VOS) method, called SHLS, that uses superpixels to build a high-compressed latent space. The proposed method is completely self-supervised, initially trained on a dataset of various orders of magnitude less than existing self-supervised VOS methods; using pseudo-labels in the offline training stage avoids the burden of annotations. The ultra-compact latent space allows for creating more efficient memory clusters, ultimately speeding up the segmentation process across the video. With efficient-oriented memory usage, SHLS achieved superior performance on single-object segmentation and comparable results with other state-of-the-art methods on multi-object segmentation on the DAVIS dataset.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

VOS aims at classifying the pixels along a frame sequence into foreground and background regions. The simplest case is single-object segmentation, where no differentiation among distinct foreground objects is required. The task becomes more challenging in the multi-object scenario, where each object in the foreground must be assigned a different label. The common approach to solve this problem is based on *supervision*. However, providing pixel-wise annotations for thousands of frames is complex, time-consuming, and costly. More recently, *self-supervised* approaches have been proposed as an alternative to allow VOS training based on completely unlabeled data. Self-supervised methods can learn inter-frame correspondences from supervisory signals extracted directly from raw videos, dispensing any level of human supervision. In principle, eliminating manual annotations is advantageous as the methods can learn from more diverse data sources. Nevertheless, many self-supervised methods have traded off the dependency on annotated frames by requiring unprecedented volumes of training videos (Fig. 1). For instance, [2], [9], [16], [33], [35], [38] are trained with enormous datasets, *e.g.*, Kinetics [3], VLOG [8], and TrackingNet [21], each one comprised of hundreds of hours of videos.

A. Originality

We introduce here a different approach by pursuing to learn VOS not only from unlabeled images but using as little training data as possible, as illustrated in Fig. 1. The proposed method, called *superfeatures in a high-compressed latent space* (SHLS), combines superpixels and deep convolutional features to produce ultra-compact representations in the superpixel domain. These representations, referred to here as *superfeatures*, are learned via a metric learning approach, in which our model clusters superfeatures when they come from parts of the same object. This process gives rise to a high-compressed latent space where correlated superfeatures compound clusters. At the inference, such clusters can be properly retrieved, identified, and used to classify the superpixels by means of a k-nearest neighbors (k-NN) algorithm. Relying on superpixels for self-supervised VOS benefits from three main aspects: (i) the lack of annotations to guide self-supervised methods in learning the object shapes makes these methods more error-prone, especially regarding the object contours;

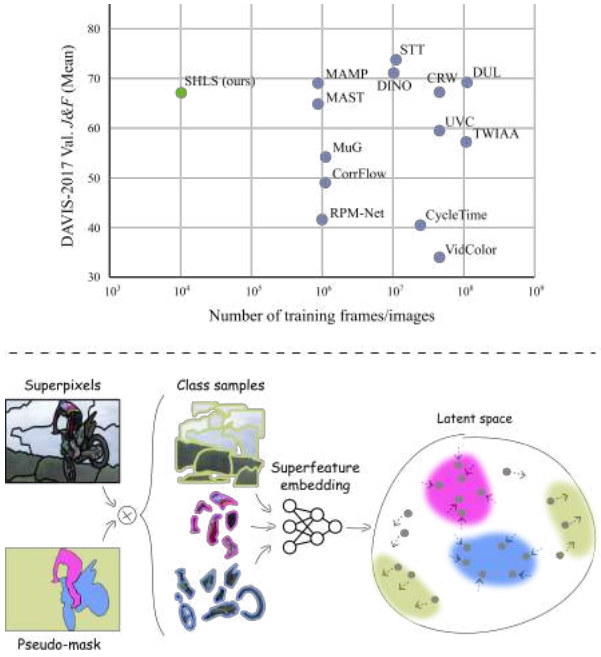

Fig. 1. The high-compressed latent space (bottom plot) generated from the superfeatures and a dataset containing at least 10^2 orders of magnitude less training images than other approaches evaluated over DAVIS-2017 (top plot).

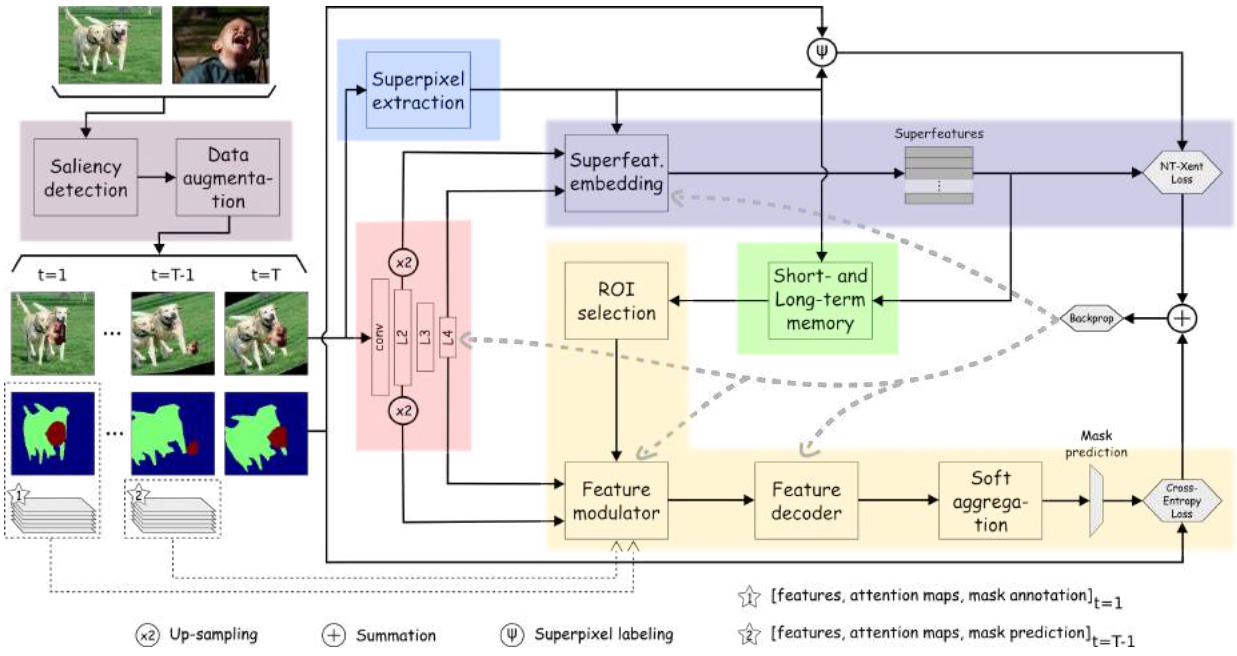

Fig. 2. Overview of SHLS at the training stage. Offline phase: Given some input still images, the pseudo-sequence generation module yields a sequence of frames and masks; following, a superpixel method extracts superpixels from the frames. Online phase: Feature maps in different scales are extracted by a CNN backbone and shared into two main branches. The uppermost branch encompasses the superpixel embedding module, which generates the superfeatures based on a contrastive NT-Xent loss. The lowermost branch accomplishes the segmentation refinement, in which the pixel-wise multi-object prediction is learned through a cross-entropy loss. This prediction is supported by the memory clustering module, which transfers information between branches by means of attention maps. At each iteration, both losses are summed and back-propagated in an end-to-end training process.

(ii) the high data compression provided by the superfeatures allows us to construct a memory mechanism that can efficiently retrieve information from virtually all past frames in a video sequence. Indeed, memory mechanisms are a crucial component for several modern VOS methods [12], [20], [24], [29], [37]; and (iii) differently from most models biased to learn features related to foreground objects only, the proposed method treats all superpixels with equal relevance, whether they come from foreground or background objects. All these aspects help SHLS to achieve more robust representations in which the background dynamics is also embedded.

To learn VOS exclusively on unlabeled data, we combine saliency detection with a set of data augmentation strategies to synthesize pseudo-sequences containing frames and masks with multi-objects. The proposed training process is based solely on the RGB images of MSRA10K [5], a relatively small dataset comprised of 10k still images. With the ground truth provided by the generated pseudo-masks, we are able to drive our superfeature embedding model toward learning a multi-class contrastive objective. The resulting superfeatures are ultra-compact vectors with dimension $1 \times S$ (in practice, we use $S = 32$), each representing the whole bunch of pixels contained in the corresponding superpixel area. Since a typical segmentation of a 480p resolution frame can be accomplished with less than a thousand superpixels, we end up with $\sim 1k \times 32$ vectors to represent each frame content – certainly a very manageable volume in computational terms. Such compactness allows us to efficiently maintain a memory clustering mechanism, where the superfeatures produced along the video processing are assigned to clusters according to the object classes present in the scene, including the background class.

Differently from methods based on retrieving information from large feature maps [14], [20], [24], [29], [37], or frames [12] accumulated in a memory bank, our superfeature mechanism does not require any special maintenance protocol to prevent overhead, and efficient similarity search [10] can access the memory. Following this approach, SHLS can learn how to carry out with VOS from a relatively small number of static images, showing competitive performance compared to state-of-the-art self-supervised methods trained with much larger datasets.

B. Impacts of the work

The dissemination of digital technologies, mainly boosted by mobile devices, has led to an exponential availability of video data. Consequently, the demand for tools to support automatic video analysis and understanding is also increasing. In this context, VOS is a crucial task, potentially benefiting areas ranging from video processing activities [31] to applications including visual tracking [25], video-based question answering [27], human pose estimation [34], surveillance [22] and so on.

C. Contributions

Our contributions include a superpixel method called Iterative Over-segmentation via Edge Clustering (ISEC) [19],

Method	Year	Training datasets		DAVIS-2016			DAVIS-2017		
		Images	Videos (hrs)	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
VidColor [33]	2018	-	K (833)	38.9	30.8	34.9	34.6	32.7	33.7
CorrFlow [13]	2019	-	O (14.0)	48.9	39.1	44.0	47.7	51.3	49.5
CycleTime [35]	2019	-	V (344)	55.8	51.1	53.5	41.9	39.4	40.7
UVC [16]	2019	-	K (833)	-	-	-	57.7	61.3	59.5
RPM-Net [11]	2020	-	D17+Y (5.75)	-	-	-	41.0	42.2	41.6
MAST [12]	2020	-	Y (5.67)	-	-	-	63.3	67.6	65.5
MUG [18]	2020	-	O (14.0)	63.1	61.8	62.5	52.6	56.1	54.3
CRW [9]	2020	-	K (833)	-	-	-	64.8	70.2	67.6
DUL [2]	2021	-	T (140)	-	-	-	67.1	71.7	69.4
TWIAA [38]	2021	-	V+K (1,177)	-	-	-	58.2	56.7	57.5
STT [15]	2022	I	Y (5.67)	-	-	-	**71.1**	**77.1**	**74.1**
MAMP [20]	2022	-	Y (5.67)	-	-	-	68.3	71.2	69.7
SHLS (ours)	2023	M	-	**76.6**	**70.4**	**73.5**	68.3	68.7	68.5

TABLE I

COMPARISON OF OUR SHLS METHOD WITH VARIOUS SELF-SUPERVISED METHODS USING STANDARD VOS METRICS, INCLUDING REGION JACCARD SIMILARITY ($\mathcal{J}$) AND BOUNDARY F-MEASURE ($\mathcal{F}$), AS WELL AS THE MEAN OF BOTH ($\mathcal{J}\&\mathcal{F}$). THE TESTS WERE PERFORMED ON THE VALIDATION SETS OF DAVIS-2016 [26] AND DAVIS-2017 [28] FOR THE SINGLE AND MULTI-OBJECT VOS TASKS, RESPECTIVELY. TRAINING DATASETS: I: IMAGENET [6]; D16: DAVIS-2016 [26]; E: ECSSD [30]; M: MSRA10K [5]; P: PASCAL-VOC [7]; D17: DAVIS-2017 [28]; Y: YOUTUBE-VOS [36]; C: COCO [17]; K: KINETICS [3]; O: OXUVA [32]; V: VLOG [8]; T: TRACKINGNET [21].

whose main characteristic is the ability to generate an adaptive number of superpixels based on the image content. This aspect is especially useful for video segmentation since the number of generated superpixels can respond to changes in the video content along the sequence.

In addition to ISEC, the main contribution of this work is summarized in SHLS. The proposed VOS method comprises several innovative characteristics, including a model based on compressed features using superpixels and metric learning, a memory mechanism based on clustering, and a new strategy to provide pseudo-labels for synthetic videos generated automatically. As far as we know, there is no related work with the same characteristics as SHLS

II. METHOD

Our SHLS framework is turned to the one-shot VOS modality. During inference, it receives the ground truth mask of the first frame and propagates it to the subsequent frames. To emulate this scenario in training, an initial offline stage is firstly accomplished, where the necessary training inputs are generated based on a bunch of still images randomly selected from the dataset [5]. Fig. 2 shows an overview of our framework. Offline-generated training inputs consist of a pseudo-sequence containing the frames and object masks and each frame's superpixel segmentation [1], [19]. Once generated, these inputs are processed sequentially at the online stage. The initial step is feature extraction, where convolutional feature maps of different scales are produced and shared into two main branches. The uppermost branch is dedicated to the superfeature generation. In this branch, the superpixel embedding network receives the features and superpixels of the current frame and generates the superfeatures according to a contrastive NT-Xent objective [4].

Along the frames, the generated superfeatures are stored by memory clustering. This module provides short- and long-term memory mechanisms to retrieve historical information to support the current frame segmentation. The memory clustering yields a set of object-focused attention maps, which are passed to the segmentation refinement branch (lowermost, in Fig. 2). Segmentation refinement is run at the pixel-level, for each foreground object individually. For this, the object region of interest (ROI) is selected from the attention maps and passed to the feature modulator and feature decoder modules. Both are network-based modules, where the former modulates the features of the current frame. This is accomplished according to the object ROI selected in the attention maps and the features, attention maps and mask prediction of the previous frame. The modulated features and the ROI-selected attention maps are then passed to the feature decoder module. There, they are fed into the decoder network along with previous information from the first and the last iterations. The feature decoder predicts individual masks for each object in the frame. Ultimately, these masks are joined via soft-aggregation [23] to generate the final multi-object prediction. A cross-entropy function computes the error between this prediction and the corresponding pseudo-mask. At each iteration, the NT-Xent and cross-entropy losses are summed and back-propagated in an end-to-end training process.

III. CONCLUSION

Our fully self-supervised training methodology enables training with only 10k still images. Our experiments on the DAVIS dataset (Table I) demonstrate that SHLS outperforms self-supervised methods by a large margin on the single-object DAVIS test and remains competitive on the multi-object test, despite being trained with significantly fewer data than competitors.

REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(11):2274–2282, 2012.

[2] Nikita Araslanov, Simone Schaub-Meyer, and Stefan Roth. Dense unsupervised learning for video segmentation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25308–25319. Curran Associates, Inc., 2021.

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[5] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015.

[6] 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, 2009.

[7] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. 88(2):303–338, jun 2010.

[8] David F. Fouhey, Weicheng Kuo, Alexei A. Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018.

[9] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *Advances in Neural Information Processing Systems*, 2020.

[10] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[11] Y. Kim, S. Choi, H. Lee, T. Kim, and C. Kim. Rpm-net: Robust pixel-level matching networks for self-supervised video object segmentation. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2046–2054, Los Alamitos, CA, USA, mar 2020. IEEE Computer Society.

[12] Zihang Lai, Erika Lu, and Weidi Xie. MAST: A memory-augmented self-supervised tracker. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[13] Z. Lai and W. Xie. Self-supervised learning for video correspondence flow. In *BMVC*, 2019.

[14] M. Li, L. Hu, Z. Xiong, B. Zhang, P. Pan, and D. Liu. Recurrent dynamic embedding for video object segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1322–1331, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society.

[15] Rui Li and Dong Liu. Spatial-then-temporal self-supervised learning for video correspondence, 2022.

[16] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[18] X. Lu, W. Wang, J. Shen, Y. Tai, D. J. Crandall, and S. H. Hoi. Learning video object segmentation from unlabeled videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8957–8967, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[19] Marcelo Mendonça and Luciano Oliveira. Isec: Iterative over-segmentation via edge clustering. *Image and Vision Computing*, 80:45–57, 2018.

[20] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Self-supervised video object segmentation by motion-aware mask propagation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022.

[21] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[22] Sundaram Muthu, Ruwan Tennakoon, Tharindu Rathnayake, Reza Hoseinnezhad, David Suter, and Alireza Bab-Hadiashar. Motion segmentation of rgb-d sequences: Combining semantic and motion information using statistical inference. *IEEE Transactions on Image Processing*, 29:5557–5570, 2020.

[23] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7376–7385, 2018.

[24] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[25] Matthieu Paul, Martin Danelljan, Christoph Mayer, and Luc Van Gool. Robust visual tracking by segmentation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 571–588, Cham, 2022. Springer Nature Switzerland.

[26] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[27] AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S. Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 76–94, Cham, 2022. Springer Nature Switzerland.

[28] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[29] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12889–12898, October 2021.

[30] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):717–729, 2016.

[31] Jayesh Vaidya, Arulkumar Subramaniam, and Anurag Mittal. Co-segmentation aided two-stream architecture for video captioning. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2442–2452, 2022.

[32] Jack Valmadre, Luca Bertinetto, Joao F. Henriques, Ran Tao, Andrea Vedaldi, Arnold W.M. Smeulders, Philip H.S. Torr, and Efstratios Gavves. Long-term tracking in the wild: a benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[33] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, page 402–419, Berlin, Heidelberg, 2018. Springer-Verlag.

[34] Urs Waldmann, Jannik Bamberger, Ole Johannsen, Oliver Deussen, and Bastian Goldlücke. Improving unsupervised label propagation for pose tracking and video object segmentation. In Björn Andres, Florian Bernard, Daniel Cremers, Simone Frintrop, Bastian Goldlücke, and Ivo Ihrke, editors, *Pattern Recognition*, pages 230–245, Cham, 2022. Springer International Publishing.

[35] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019.

[36] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part V*, page 603–619, Berlin, Heidelberg, 2018. Springer-Verlag.

[37] Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. Reliable propagation-correction modulation for video object segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2946–2954, Jun. 2022.

[38] Wenjun Zhu, Jun Meng, and Li Xu. Self-supervised video object segmentation using integration-augmented attention. *Neurocomput.*, 455(C):325–339, sep 2021.