FODVid: Flow-guided Object Discovery in Videos

Anonymous CVPR submission

Paper ID 00046

Abstract

*Segmentation of objects in a video is challenging due to the nuances such as motion blurring, parallax, occlusions, changes in illumination, etc. Instead of addressing these nuances separately, we focus on building a generalizable solution that avoids overfitting to the individual intricacies. Such a solution would also help us save enormous resources involved in human annotation of video corpora. To solve Video Object Segmentation (VOS) in an unsupervised setting, we propose a new pipeline (**FODVid**) based on the idea of guiding segmentation outputs using flow-guided graph-cut and temporal consistency. Basically, we design a segmentation model incorporating intra-frame appearance and flow similarities, and inter-frame temporal continuation of the objects under consideration. We perform an extensive experimental analysis of our straightforward methodology on the standard DAVIS16 video benchmark. Though simple, our approach produces results comparable (within a range of $\sim$ 2 mIoU) to the existing top approaches in unsupervised VOS. The simplicity and effectiveness of our technique opens up new avenues for research in the video domain.*

1. Introduction and Related Work

Object segmentation, in its various forms, is a widely studied problem in computer vision [18]. The classic task finds critical applications across multiple domains, such as autonomous driving, augmented reality, human-computer interaction, video summarization etc. It is typically solved using deep neural networks trained on large annotated datasets created through enormous human efforts that take several months of focused work. Especially in the domain of videos, annotation becomes challenging as the annotator has to add segmentation labels to the individual frame of each video in the provided corpus. Moreover, training a segmentation model on one such dataset does not guarantee transfer to real-world data since dataset-specific considerations in model design overfit the model to a particular use case. These issues (cumbersome annotations and overcatering to a dataset) call attention to developing generalizable solutions that can work with minimal human supervision.

To alleviate some of the problems around supervised segmentation, methods that work on the weaker forms of human supervision were proposed. These methods function with weak supervision provided through scribbles [14, 29, 30, 32, 38] or clicks [1] or image-level tags [21, 30, 38] or even bounds [5]. Further, semi-supervised segmentation techniques [5, 7, 9, 21, 24] were proposed that attempt segmentation in a setting where only a fraction of the image datasets are human-labelled. Nonetheless, both weak-supervised and semi-supervised techniques still rely on human supervision in some form. Such supervision, however much small, becomes bulky when applied to the domain of videos. Therefore, in the present work, we focus on segmenting objects in a video without relying on any form of external supervision.

The aim of Video Object Segmentation (VOS) is to localize the most salient object(s) in a given video frame [25]. In the literature [13,20,25,36,42], this problem is generally reframed as a foreground-background separation, where the most salient object in the video *typically* forms the foreground. The current SOTA, Choudhury *et al*. [4] incorporate a quadratic component flow model to predict regions that are likely to contain simple motion patterns. Yang *et al*. [41] propose an adversarial framework wherein a generator network tries to minimize the mutual information between segments, and another inpainting network that predicts the optical flow of the segments based on context. Along the same lines as these works, we present a complementary technique that simplifies the overall segmentation pipeline without significant loss in performance.

We propose an end-to-end pipeline as follows – for a given video frame, we compute the optical flow in RGB format and process the frame in conjugation with its flow using an image encoder like DINO [3]. The image and flow features are then used to form a similarity based adjacency matrix between the frame patches. We perform graph-cut on this adjacency matrix to obtain a preliminary set of foreground-background masks for all the frames. These masks can be used as pseudo-ground truths to train a seg-

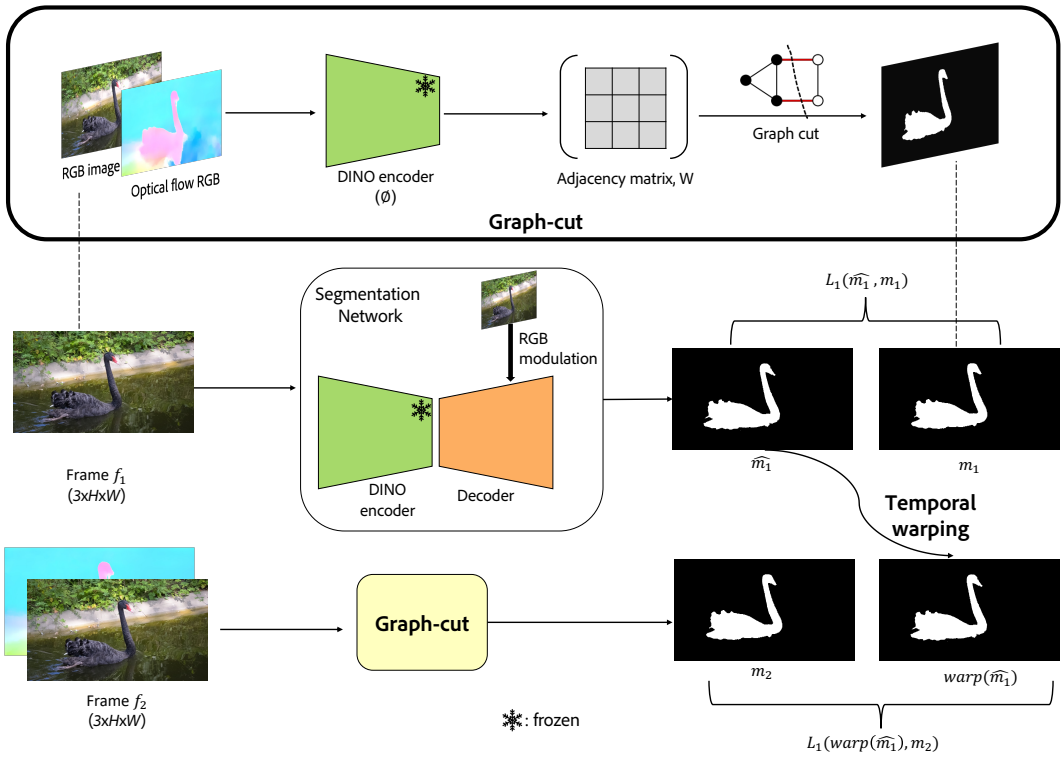

Figure 1. **FODVid: Our proposed pipeline for unsupervised video object segmentation.** An image and its corresponding optical flow RGB are featurized using DINO [3] and graph-cut is performed to produce a set of preliminary object segmentation masks. These masks are then used as pseudo ground-truths to train a segmentation network by enforcing temporal consistency between nearby frames.

mentation network. In order to enforce that the segmentation masks be temporally consistent with nearby frames' masks, we minimize the loss between warped masks and the pseudo-ground truths(refer 2.2). This methodology is straight-forward to implement and generates results of quality comparable to those of existing top methods. We summarize the contributions of this work below:

1. We present a new pipeline (FODVid) for unsupervised Video Object Segmentation (VOS) guided by optical flow. Specifically, we combine appearance and flow features to produce high-quality segmentation masks via graph-cut that are later refined using object motion information present in nearby video frames.
2. We demonstrate the importance of perceptual motion cues for object discovery. In particular, we employ the Gestalt principle, "things that move together, belong together" to enrich frame patch similarity.
3. Our methodology is simple to implement and produces results comparable to existing top unsupervised VOS approaches. We achieve an mIoU score of **78.71** on the standard DAVIS16 benchmark. Additionally, the proposed temporal refinement provides an improvement of as much as **+9.88 mIoU** on certain video sequences.

2. Methodology

Our approach involves guiding segmentation through – 1) graph-cut and 2) temporal warping via optical flow (Refer Fig. 1 for our proposed FODVid pipeline).

2.1. Graph-cut

Consider video frame f, objects of which we wish to segment out. We start by creating a fully-connected graph $G = (V, E)$, where V denotes the set of vertices obtained by dividing f into square patches of size $p_s \times p_s$, and E denotes the set of edges such that each edge weight quantifies the similarity between connecting vertices (in our case, image patches). Formally, the adjacency matrix W underlying G is made of $w_{ij} = S(v_i, v_j)$, where $S(\cdot)$ denotes the similarity measure between two given vertices (patches).

When compared to standalone images, video frames are special as they track information about a set of objects temporally, in continuation. Since the main aim of VOS is to segment out such objects, we wish to incorporate perceptual nuances in the similarity measure S. We define $S(v_i, v_j)$ based on Gestalt's principle of common fate – "things that move together, belong together". Formally, the overall similarity score between two patches from the same frame is de-

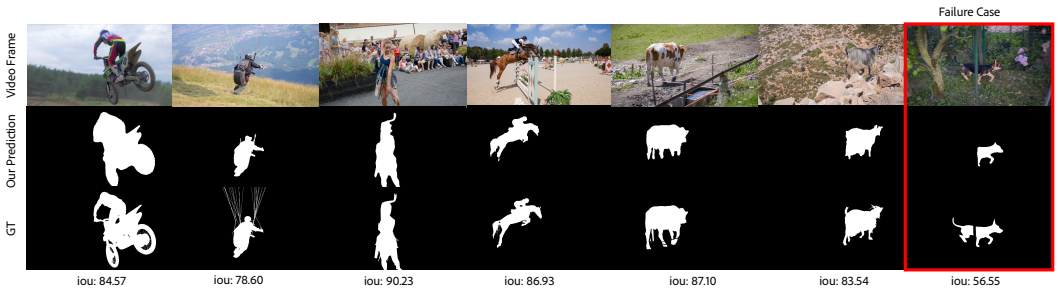

Figure 2. **Qualitative results of our method on DAVIS16 dataset**. Top row: RGB video frame, middle row: our prediction, bottom row: ground truth. The last column shows a failure case where only part of the object is identified due to occlusion.

fined as a linear combination of similarity between standard patch embeddings (obtained using DINO encoder [3]) and that between DINO embeddings of the RGB-optical flow at the respective patches. In mathematical notation,

$$S(v_i, v_j) = \alpha \cdot S'(\phi(v_i), \phi(v_j)) + (1-\alpha) \cdot S'(\phi(\psi(v_i)), \phi(\psi(v_j)))$$

where $\alpha \in [0, 1]$, $\phi(\cdot)$ denotes the DINO encoder, $\psi(\cdot)$ denotes the RGB-optical flow estimator, i.e., model computing optical flow in 3-channel RGB image format, and $S'(\cdot)$ is the cosine similarity function, given by $S'(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}||_2 ||\vec{y}||_2}$. We obtain W using the S defined above.

Further, in order to minimize the variance, we normalize w_{ij}'s by thresholding them.

$$w_{ij} \leftarrow \begin{cases} 1, & \text{if } w_{ij} \geq \tau \\ \epsilon, & \text{otherwise} \end{cases}$$

where τ denotes the weight threshold hyper-parameter, and value of ϵ is set to $10^{-5} (\neq 0)$ to ensure the fully connectedness of G.

The solution to the graph-cut is well studied in the literature [27, 34, 35]. We compute the second-smallest eigenvector of the matrix W, and threshold it to create a bi-partition of G; the partitions represent the foreground and background, respectively. We use these binary masks as ground truth for training an encoder-decoder style segmentation network.

2.2. Temporal warping via optical flow

The graph-cut approach described in the previous section relies only on information from the single frame under consideration. Although we incorporated flow-similarity, which implicitly contains object motion information, going forward, we wish to explicitly make the segmentation model discern object motion information in the video. To that end, we propose a video-level segmentation refinement scheme.

Let the frame under consideration be f_1 and the graph-cut mask obtained per the procedure described in Sec. 2.1 be m_1. We sample frame f_2 in the {-2, -1, +1, +2} temporal neighbourhood of f_1. Let m_2 denote the graph-cut mask for f_2. Let the prediction of the segmentation network for f_1 be $\hat{m_1}$.

We design a loss schedule for training the segmentation network such that, for 50% of the time, we use segmentation-loss between m_1 and $\hat{m_1}$. For the remaining 50% of the time, we temporally warp $\hat{m_1}$ using optical flow estimated between f_1 and f_2 to obtain segmentation mask prediction for f_2 and take the segmentation-loss between this mask and m_2. We use pre-trained GMFlow [37] model to estimate the optical flow in all our experiments. The algorithmic description is provided in Alg. 1.

Algorithm 1: Loss Schedule

for *epoch in* $\{1, 2, ..., N\}$ **do**
  $p \sim \mathcal{U}(0, 1)$ `// sample from uniform(0,1)`
  if $p < 0.5$ **then**
    $L = ||\hat{m_1} - m_1||_1$ `// graph-cut guidance`
  else
    $L = ||\text{warp}(\hat{m_1}) - m_2||_1$ `// enforcing temporal consistency`
  Update weights based on the computed L

Although simple, our technique produces results comparable to the existing state-of-the-art techniques. In the upcoming section, we present an extensive comparative analysis of our technique on the DAVIS16 [25] benchmark.

3. Experiments and Results

3.1. Experimental Setup

We train the segmentation network using 4 NVIDIA A100 80GB GPUs for 200 epochs with a batch size of 16. We employ Adam optimizer [11] with a learning rate of 10^{-4}, momentum terms set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The training is performed at an image resolution of 256×512. We borrow the decoder architecture from SPADE [23] and keep DINO [3] as the encoder.

Method	Flow Method	mIoU($J\uparrow$)
SAGE [33]	LDOF [2]	42.6
NLC [6]	SIFTFlow [15]	55.1
CUT [10]	LDOF [2]	55.2
FTS [22]	LDOF [2]	55.8
AMD [17]	$\times$	57.8
MG [39]	RAFT [31]	68.3
EM [19]	RAFT [31]	69.3
CIS [41]	PWCNet [28]	71.5
ARP * [12]	CPMFlow [8]	76.2
OCLR [36]	RAFT [31]	78.9
DS † [42]	RAFT [31]	79.1
DyStaB * [40]	RAFT [31]	80.0
Ponimatkin *et al.* [26]	ARFlow [16]	80.2
Choudhury *et al.* [4]	RAFT [31]	80.7
Ours (FODVid) †	GMFlow [37]	**78.71**

Table 1. **Quantitative comparison with unsupervised VOS approaches on the DAVIS16 dataset.** Our segmentation pipeline based on learning-free graph-cut combined with temporal refinement performs favourably to the existing state-of-the-art. † denotes optimization per video sequence. * DyStaB utilises supervised pre-training, ARP uses human supervision in the form of saliency maps.

Architecture	*ViT-S/8*	*ViT-S/16*	*ViT-B/8*	*ViT-B/16*
mIoU($J\uparrow$)	73.46	74.74	**76.76**	74.08

Table 2. **Ablation on the DINO architecture used for Graph-cut** (Refer Sec. 2.1). We identify ViT-B/8 as the best variant.

3.2. Results

Table 1 depicts the comparison between our method and the existing unsupervised VOS approaches. Further, in tables 2 and 3, we depict the ablations on DINO image encoder and the important hyperparameters τ (used for edge thresholding) and α (used for similarity combination). Moreover, we perform an ablation on the flow estimator used in the experiments and present its results in table 4.

We also show the qualitative results of our method on examples from the DAVIS16 dataset in figure 2. Furthermore, we demonstrate the qualitative improvements observed because of the refinement step in figure 3.

4. Conclusion

We present a novel solution to unsupervised Video Object Segmentation (VOS). Our straightforward approach is built on the idea of guiding the segmentation network with pseudo-ground truths from flow-induced graph-cut masks. We further propose adding optical flow-based temporal warping to explicitly incorporate the object motion signals in video frames. We analyse our methodology extensively on the standard DAVIS16 video dataset, where we show the effectiveness of our technique to produce results comparable to existing leading approaches. In future, we want to extend our method to other video datasets- SegTrackv2 [13] and FBMS59 [20]. We also plan to study iterative refinement of the segmentation masks through bootstrapping. Finally, through this work, we want to emphasize the importance of motion cues for object discovery.

τ	mIoU	α	mIoU
0.00	63.05	0.2	35.22
0.05	71.32	0.3	46.04
0.10	74.38	0.4	56.63
0.15	75.13	0.5	65.20
0.20	75.96	0.6	72.55
0.25	**76.76**	0.7	**75.96**
0.30	76.56	0.8	75.31
0.35	75.91	0.9	72.19
0.40	75.83	1.0	62.66

Table 3. **Hyper-parameter ablations.** We find that the similarity edge threshold $\tau = 0.25$ & the linear combination coefficient $\alpha = 0.7$ gives the best result. Further, our hypothesis around combining flow features with raw image features holds good, i.e., they improve the quality of segmentation (Refer Sec. 2.1).

Figure 3. **Improvements from guidance through temporal warping** (Refer Sec. 2.2). Top row: image, middle row: graph-cut masks, bottom row: after temporal refinement. We observe that temporal warping enables the model to predict an object when it is absent in the graph-cut mask. As shown, it generates a person's whole body while only parts of it are captured through graph-cut.

Flow Model	mIoU($J\uparrow$) on DAVIS16		
	Graph-cut	*Post Refinement*	*Post CRF*
GMFlow [37]	76.76	**77.94**	76.89
ARFlow [16]	78.03	**78.71**	77.57

Table 4. **Ablation on flow model used in Graph-cut.** Interestingly, we find ARFlow, which is trained in a completely unsupervised fashion, to perform better than its supervised alternative. Also, we observe that CRF does not improve the overall quality of obtained masks.

References

[1] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 549–565. Springer, 2016. 1

[2] Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):500–513, 2010. 4

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1, 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. *arXiv preprint arXiv:2205.07844*, 2022. 1, 4

[5] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015. 1

[6] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, volume 2, page 8, 2014. 4

[7] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. *Advances in neural information processing systems*, 28, 2015. 1

[8] Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5704–5712, 2016. 4

[9] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. 1

[10] Margret Keuper, Bjoern Andres, and Thomas Brox. Motion trajectory segmentation via minimum cost multicuts. In *Proceedings of the IEEE international conference on computer vision*, pages 3271–3279, 2015. 4

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[12] Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation and reduction. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 7417–7425. IEEE, 2017. 4

[13] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE international conference on computer vision*, pages 2192–2199, 2013. 1, 4

[14] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 1

[15] Ce Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009. 4

[16] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6489–6498, 2020. 4

[17] Runtao Liu, Zhirong Wu, Stella Yu, and Stephen Lin. The emergence of objectness: Learning zero-shot segmentation from videos. *Advances in Neural Information Processing Systems*, 34:13137–13152, 2021. 4

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[19] Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. Em-driven unsupervised learning for efficient motion segmentation. *arXiv preprint arXiv:2201.02074*, 2022. 4

[20] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200, 2013. 1, 4

[21] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 1

[22] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE international conference on computer vision*, pages 1777–1784, 2013. 4

[23] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 3

[24] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015. 1

[25] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 1, 3

[26] Georgy Ponimatkin, Nermin Samet, Yang Xiao, Yuming Du, Renaud Marlet, and Vincent Lepetit. A simple and powerful

global optimization for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5892–5903, 2023. 4

[27] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 3

[28] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 4

[29] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1818–1827, 2018. 1

[30] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018. 1

[31] 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. 4

[32] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017. 1

[33] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):20–33, 2017. 4

[34] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. *arXiv preprint arXiv:2301.11320*, 2023. 3

[35] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 3

[36] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In *Advances in Neural Information Processing Systems*, 2022. 1, 4

[37] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022. 3, 4

[38] Jia Xu, Alexander G Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3781–3790, 2015. 1

[39] 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. 4

[40] Yanchao Yang, Brian Lai, and Stefano Soatto. Dystab: Unsupervised object segmentation via dynamic-static bootstrapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2826–2836, 2021. 4

[41] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2019. 1, 4

[42] Vickie Ye, Zhengqi Li, Richard Tucker, Angjoo Kanazawa, and Noah Snavely. Deformable sprites for unsupervised video decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2666, 2022. 1, 4