Unsupervised video segmentation

Magdy Mahmoud
Technical University of Munich
Munich, Germany
magdy.mahmoud@tum.de

Shirsha Bose
Technical University of Munich
Munich, Germany
shirsha.bose@tum.de

Lisa Podszun
Technical University of Munich
Munich, Germany
lisa.podszun@tum.de

Abstract

Segmentation of the main object in videos is a fundamental problem in computer vision with broad applications in autonomous driving, robotics, surveillance, and much more. This work explores the integration of optical flow into video object segmentation, building upon the limitations of existing methods. Our work combine ideas from state-of-the-art papers, empathizing the usage of optical flow by introducing a novel consistency loss based on optical flow warping. Experiments on benchmark datasets, such as SegTrackV2, show potential improvements in video object segmentation. We show experimental results and identifies potential areas for further improvement.

1. Introduction

Segmentation of the main objects in a video is a fundamental and challenging problem in computer vision, with profound implications for a wide range of applications including robotics, autonomous driving, surveillance, social media, augmented reality and many more. Numerous methods have been proposed and explored in the literature, starting from traditional computer vision and machine learning techniques such as Saliency-Aware Video Object Segmentation [23] to more recent state-of-the-art approaches using deep learning, such as Treating Motion as Option with Output Selection for Unsupervised Video Object Segmentation [5].

Supervised video segmentation has witnessed significant progress over the years, for example, Learning Video Object Segmentation from Static Images [7] which introduced the use of convolutional neural networks (CNNs) for video segmentation had an impressive result. However, these papers rely on having large annotated datasets which are very expensive and time-consuming to annotate.

To overcome this limitation, unsupervised video segmentation witnessed big attention. Methods such as [13, 24, 26] have significantly improved the result in the unsupervised setting. However, due to the lack of the temporal information of the objects in motion these models fail to connect the objects in different time frames and thus result in poor performance of video object segmentation.

One of the tools that greatly improved the performance of unsupervised video segmentation is optical flow. Optical flow is a fundamental concept in computer vision and a powerful tool for characterizing the motion of objects within a video sequence by providing a dense representation of pixel motion between consecutive frames offering dynamic insights into motion patterns within video sequences. The recent advanced deep learning methods which are successful in achieving accurate optical flow features are [10,14,15,20,21]. Among these methods that are widely used in the deep learning literature for usage of the optical flow features are Recurrent All-Pairs Field Transforms (RAFT) model [21] and Augmentation as Regularization Flow (ARFlow) model [10].

The works that explored the utilization of the optical flow-based features for video object segmentation are [4, 6, 9, 11, 18, 22, 25]. Among them the notable methods that utilized the optical flow features in an unsupervised setting are Object-Centric Layered Representation (OCLR) [25], Guess What Moves (GWM) [6], and Semi-Supervised Learning based Video Object Segmentation (SSL-VOS) [18]. These methods share a common ideology of guiding the segmentation of the objects of interest through their motion information by utilizing the optical flow features.

Our work builds upon the recent work of SSL-VOS and GWM by combining ideas from both of them. By analyzing the failure cases of these method, we note that GWM fails

to be consistent as it's an online method that rely on one frame only to predict the segmentation without any information from adjacent frames. On the other hand SSL-VOS perform a global optimization over the whole video frames together, so we can see more consistency between adjacent frames but the segmentation masks are not good.
We propose to use the GWM network to get the powerful segmentation network, and instead of using one frame at time, we process at least two adjacent frames so that we can leverage the optical flow between them and define a consistency loss between them.

2. Related work

Our work aims to combine ideas from state-of-the-art methods, in this section, we discuss the relevant work of recent unsupervised video segmentation and optical flow methods. Then, in the next section, we study the failure cases for some of these methods.

2.1. Unsupervised video segmentation

One of the recent stat-of-the-art method in the unsupervised video segmentation setting is OCLR. The authors of OCLR [25] introduce an Object-Centric Layered Representation (OCLR) model for discovering and segmenting multiple moving objects and inferring their mutual occlusions from optical flow alone. The architecture of OCLR consists of a CNN backbone, a transformer encoder, and a DETR [1] based decoder followed a layer ordering to perform amodal segmentation. The model is trained in a supervised manner over synthetic data and is finally implemented over the real world data.
On the other hand, SSL-VOS [18] propose a novel objective based on spectral clustering for efficient video object segmentation. The method relies on appearances based features from models like DINO [2] etc, and optical flow features like RAFT [21], Arflow [10], etc. With these appearance features and optical flow features the authors construct an affinity matrix which is used for object segmentation from videos, after a global optimization of the proposed objective function.
GWM [6] takes a unique approach by combining motion-based and appearance-based segmentation. It supervises an image segmentation network by predicting regions likely to contain simple motion patterns, indicative of objects. This supervision is accomplished by estimating the motion of objects in a video. In GWM, optical flow plays a role as indirect supervision, guiding the model in identifying regions with distinct motion characteristics.
All of these methods aim to perform video object segmentation, but they diverge in how they incorporate optical flow features. OCLR directly uses optical flow as input to the network, while GWM employs optical flow as indirect supervision. In SSL-VOS, optical flow features are integrated into an affinity matrix, contributing to object segmentation decisions across different temporal frames. These nuanced differences highlight the versatility of optical flow in guiding various aspects of the video object segmentation process.

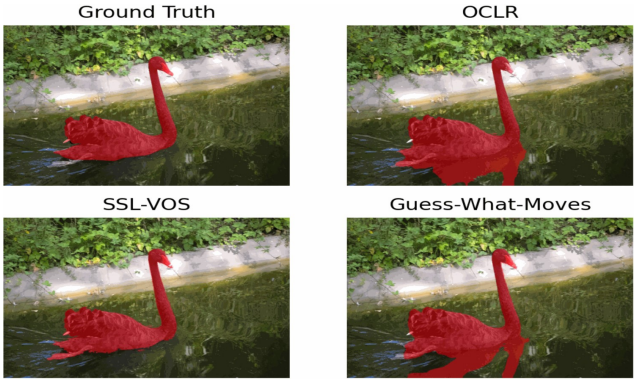

Figure 1. Failure case of reflection problem

2.2. Optical flow

One of the standard methods for optical flow estimation is RAFT [21], which is an advanced deep learning model for optical flow generation. It operates by extracting features on a per-pixel basis, constructing multi-scale 4D correlation volumes for all pixel pairs, and continually refining the flow field through an iterative process. This refinement is achieved through a recurrent unit that leverages the correlation volumes to perform look-ups and update the optical flow information.
Another fundamental method inspired by classical energy-based optical flow methods is UnFlow [16]. The authors first compute bi-directional optical flow by performing a second pass with the two input images exchanged. Then, a loss function is designed to leverage the bidirectional flow to explicitly reason about occlusion and make use of the census transform to increase robustness on real images.

3. Failure cases of state-of-the-art methods

We start our work by analyzing the failure cases for some state-of-the-art methods. One of the main failure case we noticed in OCLR and GWM is that the method confuses the main object with its shadow or its reflection, and predict them as part of the object. For example, Figure 1 show one case of this, where OCLR and GWM predict the reflection of the blackswan on the water as part of the segmentation.

We also observed a case where SSL-VOS confuses the main object with different instances of the same object. For instance, when we have several people near each other, the method fails to capture the main object as shown in Figure 2. We believe this mainly due to the fact that SSL-VOS uses clustering on the extracted DINO features, which are very

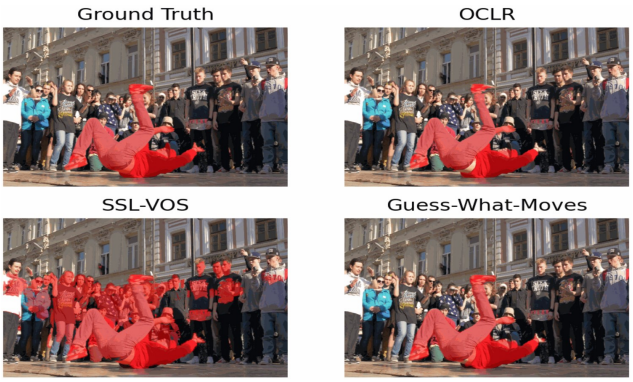

Figure 2. Failure case of multiple instances of the same object

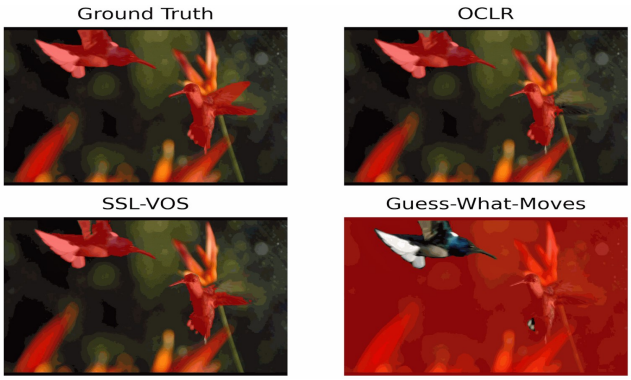

Figure 4. Failure case of flipping problem

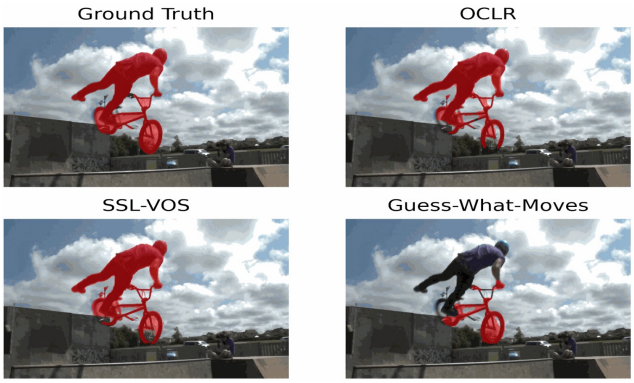

Figure 3. Failure case of consistency problem

Figure 5. Result of warping mask with optical flow

similar for different instances of the same object.
In both of these cases, the result of GWM was really good, as it uses a powerful segmentation network that can easily differentiate the main object. However, the problem with GWM is that it's not consistent across adjacent frames as it's an online method. We can see that clearly in Figure 3, where the previous frame had a correct segmentation of the person but then it misses it completely on the shown frame. Another problem we observed with GWM is that, the network itself doesn't learn to differentiate the background from the foreground, and it uses the ground truth to determine which one is which. And without using the ground truth, we have many inconsistencies between adjacent frames. Figure 4 shows an example of that, even though we have a reasonable segmentation of the main object, the method predict the background as the foreground.

4. Technical Section

4.1. Overview

Our work explores how to use optical flow to improve the unsupervised video segmentation. We begin by introducing important techniques used in the literature to use optical flows in different tasks. Then, we proceed to elaborate on the particulars of our method.

4.2. Segmentation Mask refinement

With the goal of including the global optimization approach from SSL-VOS to improve upon the GWM approach, we first experimented on the final predictions of GWM to see the effect of using the optical flow warping function from SSL-VOS. The idea was to fix large holes in the segmentation of the object that most often occurred in videos with rapid movement. Our assumption was that optical flow from one frame to another would capture the movement and therefore warping the mask from t to $t+1$ could replace the missing segmentation pixels. As Figure 5 shows, simply warping the mask does indeed enlarge the segmentation mask, but in a way that the segmentation mask is doubled and one is on top the other slightly shifted since we do not account for occlusions in the optical flow.
Occlusions in the optical flow are discrepancies of present pixels in two frames. Depending on the movement and the environment, pixels can vanish or appear between two frames, therefore becoming occluded in either forwards or backwards optical flow. The following section describes our

additional approaches on improving upon the final GWM predictions using the optical flow warping function while minding occlusions in the optical flow.

4.2.1 Optical flow occlusions

The criteria we used to determine if a pixel becomes occluded is in accordance with the method of the UnFlow paper [16]. An important change to note is that we switch the semantic of 1 and 0 for a pixel being occluded, the reason becomes evident with the following steps. For a pixel in the occlusion mask o_x^f in the forward direction, it is evaluated as 1, therefore considered as not occluded, whenever the constraint

$$|w^f(x) + w^b(x + w^f(x))|^2 < \alpha_1 \left(|w^f(x)|^2 + |w^b(x + w^f(x))|^2 + \alpha_2\right) \quad (1)$$

holds and 0 otherwise.

4.2.2 Initial refinement

In a first attempt to improve the mask estimates using the warping function and optical flow we tried to improve the mask estimates in the following manner, referring to it as initial refinement. Let $\text{mask}_t^{\text{gwm}}$ be the mask estimate from the GWM pipeline, o_t^f, o_t^b the occlusion mask at timestep t in the forwards and backwards direction respectively and w^f, w^b the forwards and backwards warping function in accordance with the SSL-VOS method. Initial refinement describes to the following method:

$$\text{refinedMask}_{t+1} = w^f(o_t^f \wedge \text{mask}_t^{\text{gwm}}) \vee \text{mask}_{t+1}^{\text{gwm}} \quad (2)$$

4.2.3 Weighted average refinement

The approach in our initial refinement has an obvious flaw which is that it only can add segmentation pixels, never delete them for an improved segmentation mask. This way however we do not use the complete information of previous and following segmentation masks. Therefore, we tried a weighted average of pixels for each pixel, determining with the previous masks $t - gap$ and the following $t + gap$ if at a given pixel x it is part of the segmentation mask or not. This way we improve the initial refinement method to adjust the segmentation mask in both ways: adding and removing pixels of the segmentation mask.
Propagated masks for each gap $\in$ gaps $= \{1, 2, 3, \dots\}$:

$$\text{prop}_{t-gap\to t}^f = w^f(o_{t-gap}^f \wedge \text{mask}_{t-gap}^{\text{gwm}})$$
$$\text{prop}_{t+gap\to t}^b = w^b(o_{t+gap}^b \wedge \text{mask}_{t+gap}^{\text{gwm}}) \quad (3)$$

Occlusion masks:

$$o_{new}^f = w^f(o_{t-1}^f) \quad (4)$$
$$o_{new}^b = w^b(o_{t+1}^b) \quad (5)$$

Weighted average for each pixel x:

$$\text{weightedSum}_x = \sum_i^{\max(gaps)} \frac{1}{2^i}(prop_{t\to gap}^f(x) + prop_{t+gap\to t}^b)$$
$$\text{normalizer}_x = \sum_j^{\max(gaps)} \frac{1}{2^i}(o_{new}^f(x) + o_{new}^b) \quad (6)$$

Pixel x for timestep t holds either 0 or 1 according to:

$$x_t = \begin{cases} 1 & \text{if } \frac{weightedSum_x}{normalizer_x} > 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

4.3. Method

4.3.1 Method Overview

The core of our method is to refine the unsupervised video segmentation by using optical flow. To exploit this idea, we modify the GWM pipeline to take two adjacent frames instead of one, to leverage the optical flow between them. Figure 6 shows our overall pipeline, by utilizing the segmentation network from GWM and modifying it to take an input of two frames, t and $t + 1$, we can make use of the optical flow warping function to define a consistency loss between the two frames to refine the predictions. We discuss this in more details below.

4.3.2 Consistency Loss

Intuitively, segmentations of two consecutive frames, t and $t + 1$ should be close to each other, and mask estimate of frame t should map to the mask estimate of frame $t + 1$ by warping the forward flow on the first frame t. Vice versa, frame mask estimate of $t + 1$ should map into the mask estimate of t by warping the backward flow on the second frame $t + 1$.
Consequently, we define a forward and backward loss to account for difference between the original prediction and estimate obtained by using optical flow warping as follows:

$$\mathcal{L}_{fwd} = CE(w^f(\text{mask}_t^{\text{gwm}}), \text{mask}_{t+1}^{\text{gwm}}))$$

$$\mathcal{L}_{bwd} = CE(w^b(\text{mask}_{t+1}^{\text{gwm}}), \text{mask}_t^{\text{gwm}}))$$

Where $CE(.)$ denotes the cross-entropy. Finally, we define our consistency loss as the sum of the forward and backward loss:

$$\mathcal{L}_{consistency} = \mathcal{L}_{fwd} + \mathcal{L}_{bwd}$$

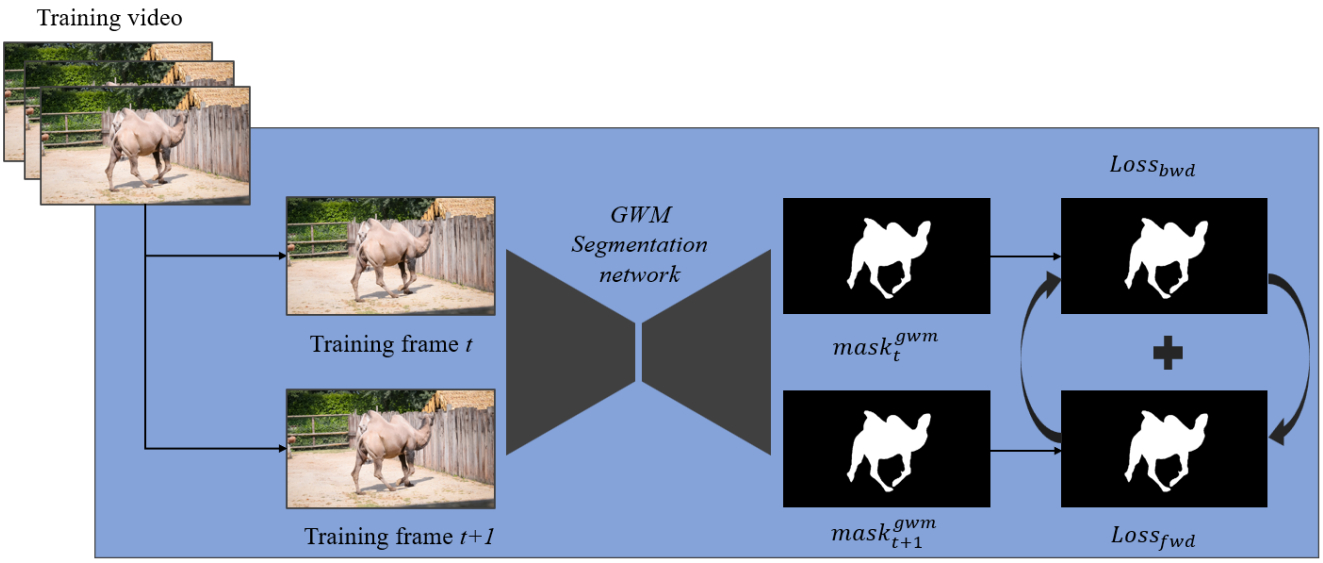

Figure 6. Our main method

5. Experiments

To evaluate our method, we conduct experiments in the setting of unsupervised video segmentation. The task is to track the main object in each frame of the video, given no annotated frames at all. First, we describe the datasets we use for comparison with other methods. Then we discuss our experiments and and provide a comprehensive analysis of the results.

5.1. Datasets

We compare our method and state-of-the-art methods on two standard benchmarks: DAVIS2016 [17], and SegTrackV2 [8]. DAVIS2016 is a densely annotated video object segmentation dataset, featuring 50 sequences and 3455 frames in total, captured in 1080p resolution at 24 FPS with precise annotation of a primary moving object at 480p. SegTrackV2 is a densely annotated dataset of 14 sequences with 976 frames in total. Some videos feature multiple objects and they have multiple challenges such as motion blur, deformations, interactions, and objects being stat.

5.2. Applying optical flow with occlusion masks on output of GWM

To start, we evaluated the usage of warping optical flows on the GWM predictions, taking into account occlusions to refine masks. For this we used the methods previously described in Section 4.2. Table 1 shows the average IoU achieved using the initial refinement method and the weighted average method using the previous and next frame and also using the previous two frames and the next two for the final mask refinement, indicated by the ranges accordingly. From the table it becomes clear that our approaches do not improve the segmentation masks, with a substantial decline with the initial refinement and the weighted average of four images.

As briefly mentioned in 4.2, the initial refinement method we used has the flaw that it is unable to decrease segmentation mask pixels. It can only increase the number of pixels and does this exceeding the actual need to improve the mask. Even with the optimal parameters α_1 and α_2 found through grid search, the large increase of segmentation mask cannot be mitigated. Since this method overshoots, it worsens the IoU score from an average of 79.5% over the DAVIS dataset to 72.3%, instead of improving it.

With this flaw in mind we tried a weighted average of multiple mask estimates, taking previous and following frames into account. However the results suggest that it does not improve the segmentation results either. On further inspection it becomes evident why this could be the case. Our method uses a weighted average over the original GWM mask estimate, the previous and following mask estimates. Self-occlusions in the occlusion masks, rapid movement and motion around the object's own axis lead the weighted average to create holes in the segmentation mask, therefore worsening the initial mask estimate. As can be seen in Table 1, the more images are taken into account the worse the effect on the segmentation mask. In this case, even if the object is moving rather slowly, as in the example of the blackswan 5, the weighted average takes too many previous and following frames into account such that the overall mask is stretched in both directions. Therefore, in all our approaches we overshoot our goal to fix missing sections in the segmentation of the main object which result in an

overall worse score.

5.3. Applying optical flow warping iteratively

To extend the first experiment for our approach, we evaluate optical flows warping on GWM prediction iteratively. That is, we start with the binary mask output from the GWM method (here we use the output from the network before the spectral clustering step), and then we use the optical flow warping function to refine the current prediction. In this experiment, we use the forward flow on previous frame and backward flow from next frame, to refine the current frame. So, for the three frames $\text{mask}_{i-1}^{\text{gwm}}$, $\text{mask}_{i}^{\text{gwm}}$ and $\text{mask}_{i+1}^{\text{gwm}}$, we update $\text{mask}_{i}^{\text{gwm}}$ as follow:

$$\alpha \cdot \text{mask}_i^{\text{gwm}} + (1-\alpha) \cdot (w^f(\text{mask}_{i-1}^{\text{gwm}}) + w^b(\text{mask}_{i+1}^{\text{gwm}}))$$

As we can see in Figure 7, the result does seem to get better after 1 iteration, but after that, a lot of noise start to spread around the object. We also noticed similar result on our other experiment 5.4, and we will be discussing this problem in more details in subsection 5.5.

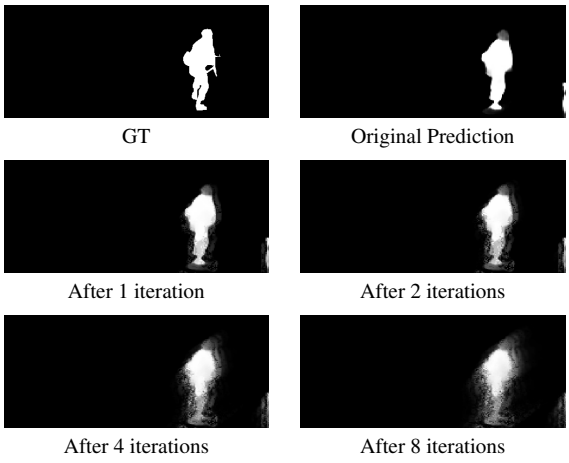

Figure 7. GWM prediction refining with optical flow

5.4. Training GWM with Consistency Loss

The main experiment for our method as discussed in 4.3, is to train the GWM network with two images at a time and define a consistency loss between them. That is, we feed two images to the network and get two predictions. We then use the optical flow between the two frames to refine the predictions.

Implementation details. We mainly use the setup of GWM. That is, we use MaskFormer [3] decoder with minor modifications to output same resolution as input. While it leverages a ViT-8 transformer, pretrained on ImageNet [19] in a self-supervised manner using DINO [2] to avoid any external sources of supervision. The networks are optimised using AdamW [12].

As mentioned before, GWM original network output 4 masks and perform spectral clustering as a post-processing in their original setup to get the foreground and background. In order to fine-tune the network with our method, we train a new model that output only 2 masks with their default configurations, that is using a learning rate $1.5 \cdot 10^{-4}$ with a schedule of linear warm-up and polynomial decay afterwards.

Then for our experiment, we use two images in a forward pass to define our loss on the output of the two images. We did many experiments with different settings, the best result we got is by freezing the backbone and only fine-tuning the decoder while continuing to use GWM optimizer and learning rate setup.

Result discussion. To investigate the impact of our approach, we discuss our method result on the bmx video from SegTrackV2. Figure 8 shows the IoU evaluation curve for this video. We get an improvement from 63 to 72 in about 500 iterations. However, the IoU start dropping very badly after about 600 iterations.

To investigate this further, Figure 5.4 shows how our method update the prediction for frame 16 from this video over different iterations. As we can see, in the original prediction, the person segmentation is almost missing, but after 80 iterations, we start to refine the head of the person in our prediction. After about 550 iterations, the person segmentation is almost there, but we start to see a lot of noise around the object. After 700 iterations, we can see that we're using the boundaries of the object segmentation, and the noise start to spread completely. This match our result from experiment 5.3, and we are going to discuss this problem in the next section.

For qualitative result, we evaluate our experiment on SegTrackV2. Our baseline here is the GWM model that we trained on SegTrackV2 that output 2 masks only, this baseline achieve an evaluation IoU of 63.6 on the SegTrackV2 dataset. Table 2 shows the result of our method on several videos separately and the sum over all of them. Overall, we didn't get much improvement over the whole dataset, but for some videos we get around 10% improvements, which shows the potential of our method.

5.5. Analysis and future work

In this section, we discuss the problems we observed with our method and suggest next steps for how we think it can be improved.

Failure analysis. As we saw from experiment 5.3 and experiment 5.4, we have a big issue of noise spreading around the main object segmentation. As previously stated, we follow the work of SSL-VOS and GWM, so we use the warping function and pre-computed optical flows from SSL-VOS. In the SSL-VOS paper, they do some post-processing

	Initial refinement	Weighted average (-1,+1)	Weighted average (-2, +2)	GWM
DAVIS	0.7129	0.7856	0.6279	79.5

Table 1. Average IoU for each refinement method used on the listed datasets, compared with the IoU of GWM

Sequence	GWM (2 masks)	Ours
Soldier	72.4	75.9
Hummingbird	62.7	71.6
Bmx	64.2	74.2
All	63.6	63.8

Table 2. Evaluation IoU on SegTrackV2 videos

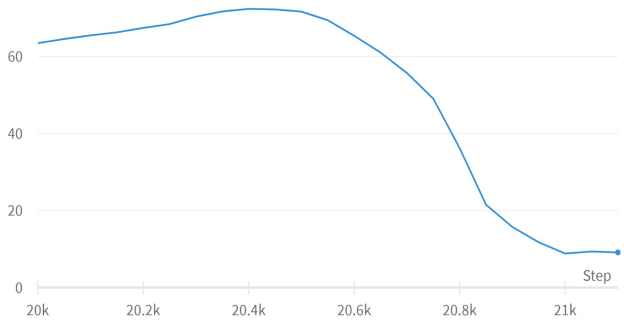

Figure 8. IoU evaluation result for the bmx video from SegTrackV2 using our method

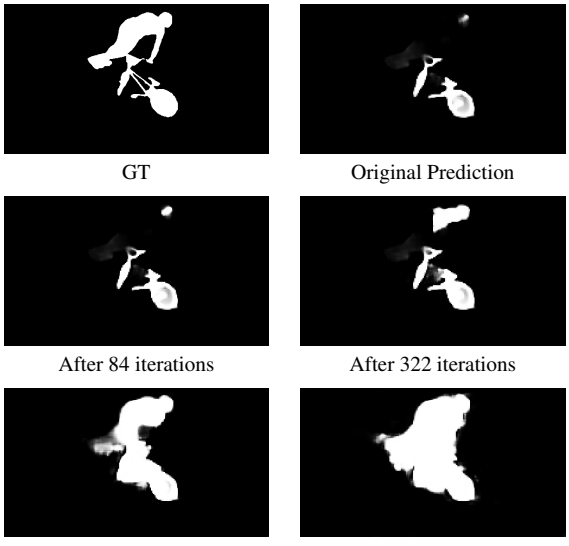

Figure 9. Visualization of the main method prediction on one frame over different iterations

to filter the wrong or poor quality optical flows predictions (filtering the locations with high response on the difference between original image, and the image reconstructed with optical flow). While investigating our noise issue, it

Figure 10. SSL-VOS pre-computed optical flow

turns out that after this filtering, a lot of noise is introduced around the main object, as we can see in Figure 10 and we didn't account for this in our method formulation. More importantly, in the SSL-VOS method, they add another loss to regularize the deviation between the method segmentation prediction and the initial prediction, and add a constant weight to balance the effect of both losses (difference from initial segmentation vs flow discrepancy).
In our work, we didn't add this deviation loss, as we thought using the pre-trained GWM and fine-tuning the decoder only would be enough to not deviate too much from initial segmentation, but as the analysis of our approach suggest, we clearly need to account for that.
Future work. We believe regularizing the deviation from initial segmentation would be a very important step to alleviate the issue of spreading noise that we have. Also, as a next step, we want to experiment with mask refinement using occlusion mask as we did in experiment 4.2. Another important step is to train GWM from the beginning while using the consistency loss with some regularization.
Another important direction is to change the final non-differentiable step of spectral clustering into a differentiable clustering and train the network end to end. Furthermore, as we see from our results, some segmentation masks get improved in a few iterations while some get bad very quickly, so we want to explore heuristic methods as stopping conditions to determine the number of iterations we should perform for each video.

6. Conclusion

We presented a comprehensive study of using optical flows to improve the result of unsupervised video segmen-

tation. Even though our main method didn't significantly improve the result of the current methods, our main experiment study shows the potential impact our method can achieve. However, in our refinement approach working on the GWM mask estimates, the optical flow methods we used decrease the masks' precision by either enlarging the segmentation mask too much or creating holes in it. Therefore, with these post-processing methods we could not achieve a refinement of the GWM segmentation masks using optical flow.

7. Statement of contributions

7.1. Magdy Mahmoud

Worked on verifying the result of the related work section 2. Exploring failure cases for different methods 3. Doing research for the main method, by investigating recent SOTA work and combining ideas from them and propsed our main method. Worked on creating the mid-term and final presentation. Implemented the first version of doing a post-processing optimization like SSL-VOS on top of GWM predictions. Implemented the pipeline for the main method 4.3. Implementing the main experiments for our work in section 7 and section 9. Worked on some experiments to account for occlusion from UnFlow in our main method.

7.2. Lisa Podszun

Recreated the results of related work and verifying the results 2. Explored failure cases and main problem in GWM 3. Did research in recommended work to design a main method and started implementing it. Implemented visualization and exploring refinement methods. Worked on and finalized mid-term and final presentation. Worked on refinement with occlusions and analysis 4.2.

7.3. Shirsha Bose

Worked on verifying the result of the related work sections and exploring failure cases along with implementation of the visualizations. Did research for the main method and one idea was proposed which was not adapted in this practical but kept open for future works. Doing experiments on existing methods. Worked on the introduction section and the related works section. Worked on creating the mid-term and final presentation.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[2] 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. 2, 6

[3] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *CoRR*, abs/2107.06278, 2021. 6

[4] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, pages 686–695, 2017. 1

[5] Suhwan Cho, Minhyeok Lee, Jungho Lee, MyeongAh Cho, and Sangyoun Lee. Treating motion as option with output selection for unsupervised video object segmentation, 2023. 1

[6] 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, 2

[7] Anna Khoreva, Federico Perazzi, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images, 2016. 1

[8] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg. Video segmentation by tracking many figure-ground segments. In *2013 IEEE International Conference on Computer Vision*, pages 2192–2199, 2013. 5

[9] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 207–223, 2018. 1

[10] 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. 1, 2

[11] Ziyang Liu, Jingmeng Liu, Weihai Chen, Xingming Wu, and Zhengguo Li. Faminet: Learning real-time semisupervised video object segmentation with steepest optimized optical flow. *IEEE Transactions on Instrumentation and Measurement*, 71:1–16, 2021. 1

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 6

[13] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3623–3632, 2019. 1

[14] Yawen Lu, Qifan Wang, Siqi Ma, Tong Geng, Yingjie Victor Chen, Huaijin Chen, and Dongfang Liu. Transflow: Transformer as flow learner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18063–18073, 2023. 1

[15] Kunming Luo, Chuan Wang, Shuaicheng Liu, Haoqiang Fan, Jue Wang, and Jian Sun. Upflow: Upsampling pyramid for unsupervised optical flow learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1045–1054, 2021. 1

[16] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss, 2017. 2, 4

[17] 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 (CVPR)*, June 2016. 5

[18] 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 *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5881–5892, 2023. 1, 2

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 6

[20] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1599–1610, 2023. 1

[21] 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, 2

[22] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3899–3908, 2016. 1

[23] 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, 2018. 1

[24] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3064–3074, 2019. 1

[25] 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, 2

[26] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 931–940, 2019. 1