CLVOS23: A Long Video Object Segmentation Dataset for Continual Learning

Amir Nazemi, Zeyad Moustafa, Paul Fieguth
University of Waterloo, Waterloo, Ontario, Canada
{amir.nazemi,zeyad.moustafa,paul.fieguth}@uwaterloo.ca

Abstract

Continual learning in real-world scenarios is a major challenge. A general continual learning model should have a constant memory size and no predefined task boundaries, as is the case in semi-supervised Video Object Segmentation (VOS), where continual learning challenges particularly present themselves in working on long video sequences. In this article, we first formulate the problem of semi-supervised VOS, specifically online VOS, as a continual learning problem, and then secondly provide a public VOS dataset, CLVOS23, focusing on continual learning. Finally, we propose and implement a regularization-based continual learning approach on LWL, an existing online VOS baseline, to demonstrate the efficacy of continual learning when applied to online VOS and to establish a CLVOS23 baseline. We apply the proposed baseline to the Long Videos dataset as well as to two short video VOS datasets, DAVIS16 and DAVIS17. To the best of our knowledge, this is the first time that VOS has been defined and addressed as a continual learning problem. The proposed CLVOS23 dataset has been released at https://github.com/Amir4g/CLVOS23.

1. Introduction

The goal of Video Object Segmentation (VOS) is to accurately extract a target object at the pixel level from each frame of a given video. In general, there are two categories of VOS solutions: semi-supervised or one-shot VOS, in which the ground-truth masks of the target objects are given in at least one frame at inference time, and unsupervised VOS, in which the VOS model knows nothing about the objects.

Among semi-supervised VOS approaches, online VOS approaches [5,29,37] update a part of the VOS model based on the evaluated frames and estimated masks. The idea is that videos contain relevant information beyond just the given frame's mask, which a model can exploit by learning during the evaluation process.

Online model learning, *while* a video is being analyzed, leads to questions regarding how effectively the model learns from frame to frame, particularly when some aspect of the video looks different than what had been given in the ground-truth frame. This leads to the domain of continual learning, which is a type of machine learning where a model is trained on a sequence of tasks, and is expected to continuously improve its performance on each new task while retaining its ability to perform well on previously-learned tasks.

The current state-of-the-art semi-supervised and specifically online VOS methods [5,29,37] perform well on VOS datasets with *short* videos (up to a few seconds or 100 frames in length) such as DAVIS16 [35], DAVIS17 [35], and YouTube-VOS18 [45]. However, most of these methods do not retain their expected performance on long videos, such as those in the Long Videos dataset [24] as shown in the XMem paper [10]. The question of the poor performance of online VOS on long videos has not been investigated in the VOS field, nor addressed through continual learning.

Continual learning methods are typically tested on classification datasets, like MNIST [22], CIFAR10 [21], and Imagenet [13], or on datasets specifically designed for continual learning, such as Core50 [28]. The classification dataset is fed to the model as a sequential stream of data in online continual learning methods [3]. In contrast to the aforementioned datasets and test scenarios, long video object segmentation has numerous real-world applications, such as video summarization, human-computer interaction, and autonomous vehicles [48].

In this paper, we formulate and address the inefficient performance of the online VOS approaches on long videos as an online continual learning problem. Moreover, we propose a new long-video object segmentation dataset for continual learning (CLVOS23), as a much more realistic and significantly greater challenge for testing VOS methods on long videos. As a baseline, we propose a Regularization-based (prior-focused) Continual Learning (RCL) solution to improve online VOS.

2. Related work

Semi-supervised VOS methods try to maximize the benefit from whatever information is given, normally the first frame of the video. Early solutions in the literature [6, 34] fine-tuned a pretrained VOS on the given information in a video at evaluation time. In contrast, current state-of-the-art solutions attempt to benefit from previously evaluated frames and make use of an allocated memory to preserve that information from preceding frames in segmenting the current frame. The so called memory-based VOS approaches [5, 10, 30, 33, 37, 50] also are categorised into two streams, matching-based and online:

- Matching-based VOS methods [11,18,25,27,40,44,47] match the representations of previous frames, stored in memory, with the corresponding features extracted from the current frame.
- Online VOS [5,6,26,37,42] update (fine-tune) a small model based on the features and estimated masks of preceding frames.

Continual learning [1, 17, 46] is a sequential learning process where the data sequence may come from different domains and tasks; thus, a model is learning from data where distribution drift [16] may occur suddenly or gradually. Catastrophic forgetting is the key challenge in continual learning and it was first defined on neural networks [31,36] when a neural network model is trained on a sequence of tasks, but has access to the training data for only the current task. In such circumstances, the model learning process is inclined to frequently update those parameters which are heavily influenced by data from the current task, leading to previously-learned tasks to be partially forgotten. The concept of catastrophic forgetting was also defined on other machine learning models [14]. There are three different approaches to catastrophic forgetting: prior-focused (regularization-based) [9, 12], likelihood-focused (rehearsal-based) [4, 7, 43, 49], and hybrid (ensemble) approaches [23, 39].

Elastic Weight Consolidation (EWC) [20] and Memory Aware Synopses (MAS) [2] are two examples of prior-focused methods that employ regularization during training to limit the change of previously learned weights. These methods assume that previously learned task weights can serve as a prior for the current network weights, which are in charge of learning new tasks. Through the use of a penalty term in the loss function, these methods aim to preserve the significant parameters from preceding tasks.

Likelihood-focused (rehearsal) techniques concentrate on minimizing the model's loss function by taking into account historical information. Examples include deep generative replay (DGR) [41] and variational generative replay (VGR) [15], which keep previous data or train generative models on earlier tasks prior to training the new task. Generative Adversarial Networks (GANs) are used in [41] to produce data from each task as samples to be used during the training of a new task.

Finally, as their name implies, hybrid methods seek to combine the benefits of prior-focused and likelihood-focused techniques. As an example, Variational Continual Learning (VCL) [32] combines the posterior from the previous task (i.e., the prior to the current task) with information about the new task (i.e., its likelihood).

The solution proposed in this article is a Regularization-based Continual Learning (RCL) approach, drawing its motivation from EWC [20].

3. Problem formulation

An online VOS model O_Ξ [5, 29, 37] is first trained offline to minimize the following loss function and to learn the model parameters Ξ:

$$\Xi = \arg\min_{\Xi'} \mathcal{L}(O_{\Xi'}(F), Y). \tag{1}$$

In Eq. (1), $\mathcal{L}$ is usually a pixel-wise cross entropy loss [8], F is an image frame and Y is the segmented mask in which each pixel of F is labeled, based on the number of objects in the video sequence. For example, in the case of single-object video, Y is just a binary foreground/background mask. An online VOS model typically has a U-Net encoder-decoder structure [38], and further comprises the following pieces:

1. A pretrained encoder, extracting feature X from each frame F;
2. A memory $\mathcal{M} = \{\mathcal{X}, \mathcal{Y}\}$, storing features $\mathcal{X}$ and their associated labels $\mathcal{Y}$ / masks. The memory can be updated with input feature X_t and estimated output Y_t at time t;
3. A target model C^t, which is trained on the memory $\mathcal{M}^t$ at time t, and provides information to decoder D;
4. Pretrained decoder D and label encoder E [5] networks which obtain temporal information from the target model alongside the encoder's output, to generate a fine-grain output mask Y from frame F.

The time index t is based on input time frame. Thus, at time t, $\mathrm{C}^{t-\Delta_\mathrm{C}}$ is updated to C^t on $\mathcal{M}^t$ where Δ_C is the target model update step. Next, the output Y_{t+1} is estimated from C^t, thus $\mathcal{M}^t$ can be augmented with pairs (X_{t+1}, Y_{t+1}) to create $\mathcal{M}^{t+1}$. Potentially, we could update $\mathcal{M}$ at every time frame t, but for practical and computational reasons, we can choose to update the memory every $\Delta_\mathcal{M}$ frames, where $\Delta_\mathcal{M}$ is the memory update step. An analogous target model

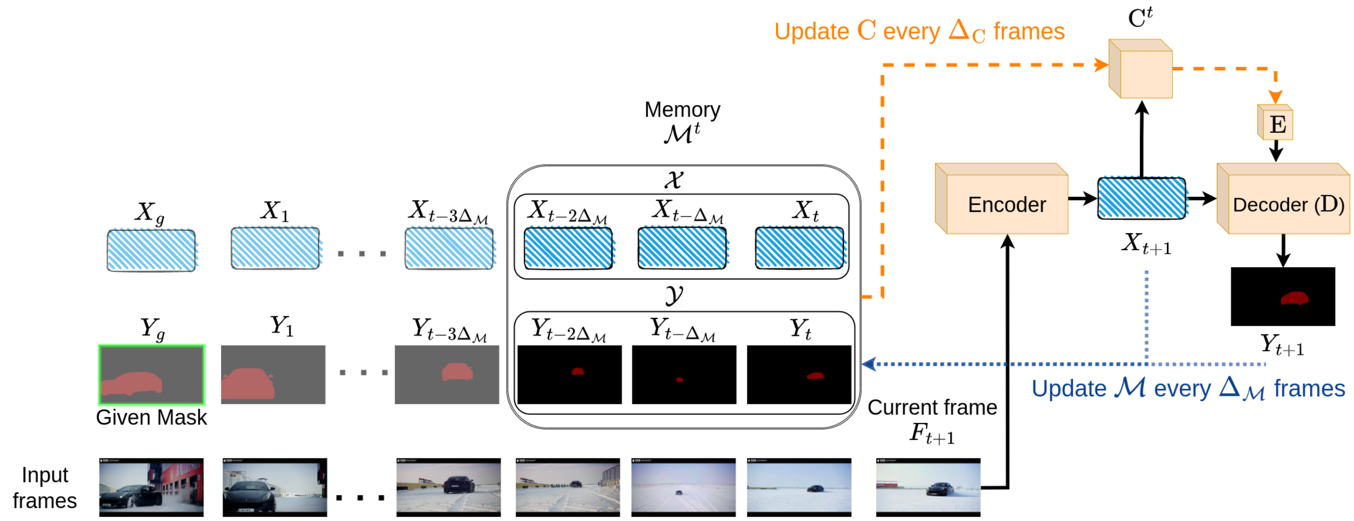
Update C every Δ_{C} frames
C^t
E
Memory
$\mathcal{M}^t$
$\mathcal{X}$
X_g
X_1
$X_{t-3\Delta_{\mathcal{M}}}$
$X_{t-2\Delta_{\mathcal{M}}}$
$X_{t-\Delta_{\mathcal{M}}}$
X_t
Encoder
X_{t+1}
Decoder (D)
$\mathcal{Y}$
Y_g
Y_1
$Y_{t-3\Delta_{\mathcal{M}}}$
$Y_{t-2\Delta_{\mathcal{M}}}$
$Y_{t-\Delta_{\mathcal{M}}}$
Y_t
Y_{t+1}
Given Mask
Update $\mathcal{M}$ every $\Delta_{\mathcal{M}}$ frames
Current frame
F_{t+1}
Input frames

Figure 1. General online VOS framework: The target model $\mathrm{C}^{\mathrm{t}-\Delta_{\mathrm{C}}}$ is updated on memory $\mathcal{M}^t$ to form C^{t}. The target model C is initialized based on the given ground truth mask Y_g and its associated feature X_g. The memory $\mathcal{M}^t$ is updated every Δ_{M} time steps (video frames) with new information (X_{t+1}, Y_{t+1}). The dashed lines show how the target model C is updated based on memory $\mathcal{M}$ every Δ_{C} frames, and the dotted lines show memory update. Our proposed methods focus on the target model component (C) of the framework. The frame images used in the figure are taken from the "car" video in the proposed CLVOS23 dataset.

update step Δ_{C} is considered for updating C. This process is depicted in Figure 1.

All of the parameters of the VOS model (Ξ) are first trained offline on a set of training data containing video frames and annotated labels; however, certain parameters of the model need to be updated online at testing time on the extracted features $\mathcal{X}$ of evaluated frames and their associated predicted labels $\mathcal{Y}$ which are kept in the memory $\mathcal{M}$. In particular, let Θ be the parameters of target model C, consisting mainly of convolutional filter weights, for $\Theta = \{\theta_l\}_{l=1}^{K}$ where K is the number of target model parameters. It should be emphasized that Θ is a rather small subset of the overall parameter set (Ξ), since the target model C is usually a small convolutional neural network for reasons of efficiency. The target model is updated every Δ_{C} frames throughout the video, repeatedly trained on features $\mathcal{X}$ and associated encoded labels $\mathrm{E}(\mathcal{Y})$ of stored decoder outputs $\mathcal{Y}$ from preceding frames. Both $\mathcal{X}$ and $\mathcal{Y}$ are stored in memory $\mathcal{M}$, as shown in Figure 1.

It is worth noting that E is a label encoder, generating sub-mask labels from each Y [5]. For online training of $\mathrm{C}^{t-\Delta_{\mathrm{C}}}$ at time t, every $Y \in \mathcal{M}^t$ is fed to E and we seek a trained model C^t to learn what E specifies from each Y. That is, the target model acts like a dynamic attention model to generate a set of score maps $\mathrm{E}\big(\mathrm{C}^t(X)\big)$ in order for the segmentation network (D) to produce the segmented output mask Y associated with each frame F. The loss function L, which is used for the online training of target model C^t at time t, is

$$L(\Theta^t, \mathcal{M}^t) = \sum_{n=1}^{|\mathcal{M}^t|} \left\| d_n W_n \Big(\mathrm{E}(Y_n) - \mathrm{E}\big(\mathrm{C}^t(X_n)\big)\Big) \right\|_2^2 + \sum_{k=1}^{K} \lambda\, {\theta_k^t}^2, \qquad (2)$$

where $\theta_k^t \in \Theta^t$ is a parameter of C^t and $|\mathcal{M}^t|$ is the number of feature and mask pairs $\{X, Y\}$ in the memory $\mathcal{M}^t$.

Depending on the overall architecture, E is an offline / pre-trained label encoder network, as in [5], or just a pass-through identity function, as in [37]. It is worth noting that the influence and effect of E is not the focus or interest of this paper.

In Eq. (2), W_n is the spatial pixel weight, deduced from Y_n, and d_n is the associated temporal weight decay coefficient. In the loss function $L(\Theta^t, \mathcal{M}^t)$, W_n balances the importance of the target and the background pixels in each frame, whereas d_n defines the temporal importance of pair of feature and mask (X_n, Y_n) in memory, typically emphasizing more recent frames [5].

4. Proposed dataset

As shown in Figure 1, online VOS assumes the change in each video sequence to be gradual, meaning that a constant size of memory $\mathcal{M}^t$ has an adequate capacity to update the target model $\mathrm{C}^{t-\Delta_{\mathrm{C}}}$ to C^t for segmenting the current frame F_{t+1}. In the ideal case, where the samples in a video sequence are independent and identically distributed (i.i.d.),

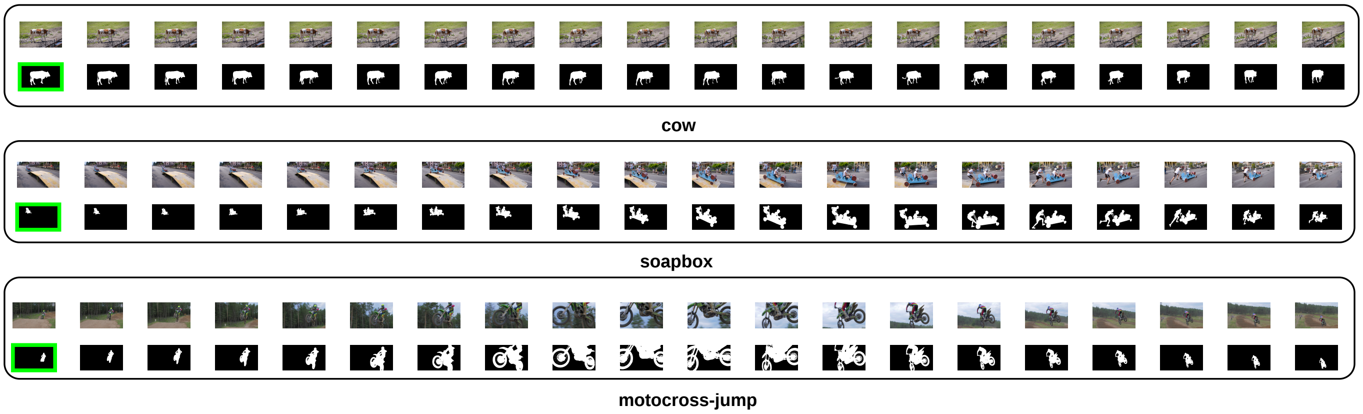

Figure 2. A set of sub-sampled frames from three videos of the DAVIS16 dataset [35], in each case two rows: actual images (top) and segmented objects (bottom). The first video, "cow" is the longest in DAVIS16, however there is no significant change between frames. There is a gradual change in appearance in the other two videos. The given annotated (ground-truth) frame in each video is highlighted in green.

machine learning problems are made significantly easier, since there is then no need to handle distributional drift and temporal dependency in VOS. However, i.i.d. assumption is not valid in video data.

Figure 2 shows three video sequences from the DAVIS2016 dataset, where we can see that target objects do not have an abrupt change through video frames. Objects could have small changes, such as in the "cow" video (the longest video in DAVIS2016 at 104 frames), and the other two videos (soapbox and motocross-jump) possess variations in object appearance, however the changes are gradual. As a result, for such datasets the identically distributed assumption of frames is usually valid, particularly for short videos. It is thus worth mentioning that the YouTube-VOS18 sequences are even shorter than those in DAVIS16 and DAVIS17, where the longest video in the validation set of YouTube-VOS18 has 36 frames.

The semi-supervised VOS approaches maintain the i.i.d. assumption for video sequences, despite the fact that this assumption is clear not valid in all video sequences, particularly longer ones. It is precisely for this reason that state-of-the-art semi-supervised VOS models are not expected to have a similar performance on long video datasets [10].

Figure 3 shows the "dressage" video from the Long Videos dataset [24], the dataset consisting of three long sequences with a total of 7411 frames. As is clear from Figure 3, an i.i.d. assumption is not at all valid on "dressage" video, because of the 22 substantial distribution drifts which take place, a behaviour which is much more closely aligned with the *non*-i.i.d. assumption of continual learning. However, this new continual learning-based interpretation of the long video sequences is discussed for the first time in VOS and continual learning. As the evaluation label mask is chosen uniformly in the Long Videos dataset, it does not show how well a VOS solution handles sudden shifts in the target's appearance. Alternatively, we propose annotating the frames for the evaluation based on the distribution drift that occurs in each video sequence.

Figure 3 shows 23 sub-chunks of videos in the "dressage" video of the Long Videos dataset. Each sub-chunk is separated from its previous and next sub-chunks based on the distribution drifts. When an online or offline event, such as a sports competition, is recorded using multiple cameras, these distribution drifts are common in media-provided videos. As a result, in our proposed dataset, we first utilize the following strategy to select candidate frames for annotation and evaluation.

- We select the first frame of each sub-chunk S. It is interesting to see how VOS models handle the distribution drift that happens in the sequence, which is arriving a new task in continual learning.
- The last frame of each sequence is also selected. The first frame ground truth label mask is given to the model as it is set in the semi-supervised VOS scenario.
- One frame from the middle of each sub-chunk is also selected for being annotated.

As shown in Figure 3, selecting the annotated frames uniformly will cause some small sub-chunks ($S_{11}, S_{12}, S_{17}, S_{19}$) to be missed in the evaluation. For CLVOS23, in addition to the 3 videos from the Long Videos dataset, we added the other 6 videos described in Table 1. All frames of the 6 new added videos are extracted with the rate of 15 Frames Per Second (FPS). To ensure that all distribution drifts are captured, we only annotate the

Video name	#Sub-chunks (tasks)	#Frames	#Annotated frames
dressage	23	3589	43
blueboy	27	1416	47
rat	22	2606	42
car	18	1109	37
dog	12	891	25
parkour	24	1578	49
skating	5	778	11
skiing	5	692	11
skiing-long	9	903	19

Table 1. Each video sequence's specifications in the proposed CLVOS23 dataset. The first three videos (Dressage, Blueboy, and Rat) are taken directly from the Long Videos dataset [24] and we added additional annotated ground-truth frames to each of them to make them more appropriate for continual learning.

first frame of each sub-chunk in the Long Videos dataset and add them to the uniformly selected annotated frames. The proposed dataset has following advantages over the Long Videos dataset [24].

- It added 5951 frames to 7411 frames of the Long Videos dataset.
- CLVOS23 increased the number of annotation frames from 63 in the Long Videos dataset to 284.
- It increases the number of videos from 3 to 9.
- The selected annotated frames are chosen based on the distribution drift that happens in the videos (sub-chunks) rather than being uniformly selected.

It is worth noting that for a long VOS dataset, it is very expensive and sometimes unnecessary to annotate all the frames of videos for evaluation. It is worth mentioning that We utilized the Toronto Annotation Suite [19] to annotate the selected frames for evaluation. The frames of new 6 videos were resized to have a height of 480 pixels. The width of each frame is defined as proportionate to its height. The link to access to the dataset is provided.[1]

5. Proposed method

A continual learning system should have a limited constant memory which is essential for a bounded system working on an infinite sequence of data. Thus, we focus on addressing continual learning using the memory-based VOS models and among them we are interested in the online VOS approaches, where part of the model (C) is updating on a constant size memory $\mathcal{M}$.

The LWL method [5], which is an extension over the well-known FRTM framework [37] benefits from a label encoder network E that tells the target model C what to learn [5]. In this article, LWL has been chosen as the online VOS baseline method. The framework structure that is explained in Figure 1 is followed by LWL, where encoder, decoder D, and the label encoder E are all trained offline; consequently, we do not make any modifications to these components by implementing the proposed solution.

The proposed regularization-based continual learning (RCL) method is inspired by the EWC [20] algorithm, where the network parameters Θ of the target model C in LWL are regularized to preserve the important parameters and prevent modification during the target model updating steps. The importance of each parameter θ_k is associated with the magnitude of its related gradient ϕ_k during the preceding update steps. Therefore, during each updating (online learning) step t, the training parameters Θ^t are regularized by the magnitude of the gradients of the target models' parameters $\Phi = \{\phi_k\}_{k=1}^{K}$ and the updated model's parameters $\Theta = \{\theta_k\}_{k=1}^{K}$ of preceding updates, which are stored in the regularizer memory $\mathcal{M}_R$.

Thus, for all features $\mathcal{X}$ and their related output masks $\mathcal{Y}$ in the memory $\mathcal{M}^t$, the target model C^t with parameters Θ^t, and the regularizer memory $\mathcal{M}_R^{t-\Delta_\mathrm{C}}$, the following loss function defined in Eq. (3) is used for training the target model of LWL:

$$L_R(\Theta^t, \mathcal{M}^t, \mathcal{M}_R^{t-\Delta_\mathrm{C}}) = \\ L(\Theta^t, \mathcal{M}^t) + \lambda \sum_{j=1}^{|\mathcal{M}_R^{t-\Delta_\mathrm{C}}|} \Phi^j \left|\left|\Theta^t - \Theta^j\right|\right|^2 \tag{3}$$

where the loss function L is described in Eq. (2), λ controls the regularisation term, and $|\mathcal{M}_R^{t-\Delta_\mathrm{C}}|$ shows how many pairs of $\{\Theta, \Phi\}$ have been stored in $\mathcal{M}_R$ so far. The loss function in Eq. (3) is used to update the target model, and it regularizes the target model training to preserve its previously learned knowledge. The proposed RCL method is depicted in Figure 4. As illustrated in this figure, the proposed RCL can be added to any online VOS method and improve its performance as shown in Section 6.

It is worth noting that the memory $\mathcal{M}$ is initialized by the encoded features of the given frame F_g and its provided ground-truth mask Y_g as defined in a semi-supervised VOS scenario.

One drawback of the proposed regularization-based method is that it needs to store the parameter importance Φ^t and the parameters of the target model Θ^t after each online updating step t; however, a limited number of stored pairs of $\{\Phi, \Theta\}$ are enough to regularize the updating step of the target model C^t.

Additionally, for a small target model C, it is feasible to calculate and store the Φ and Θ during the updating step; however, it is a real challenge for a larger target model.

[1] https://github.com/Amir4g/CLVOS23

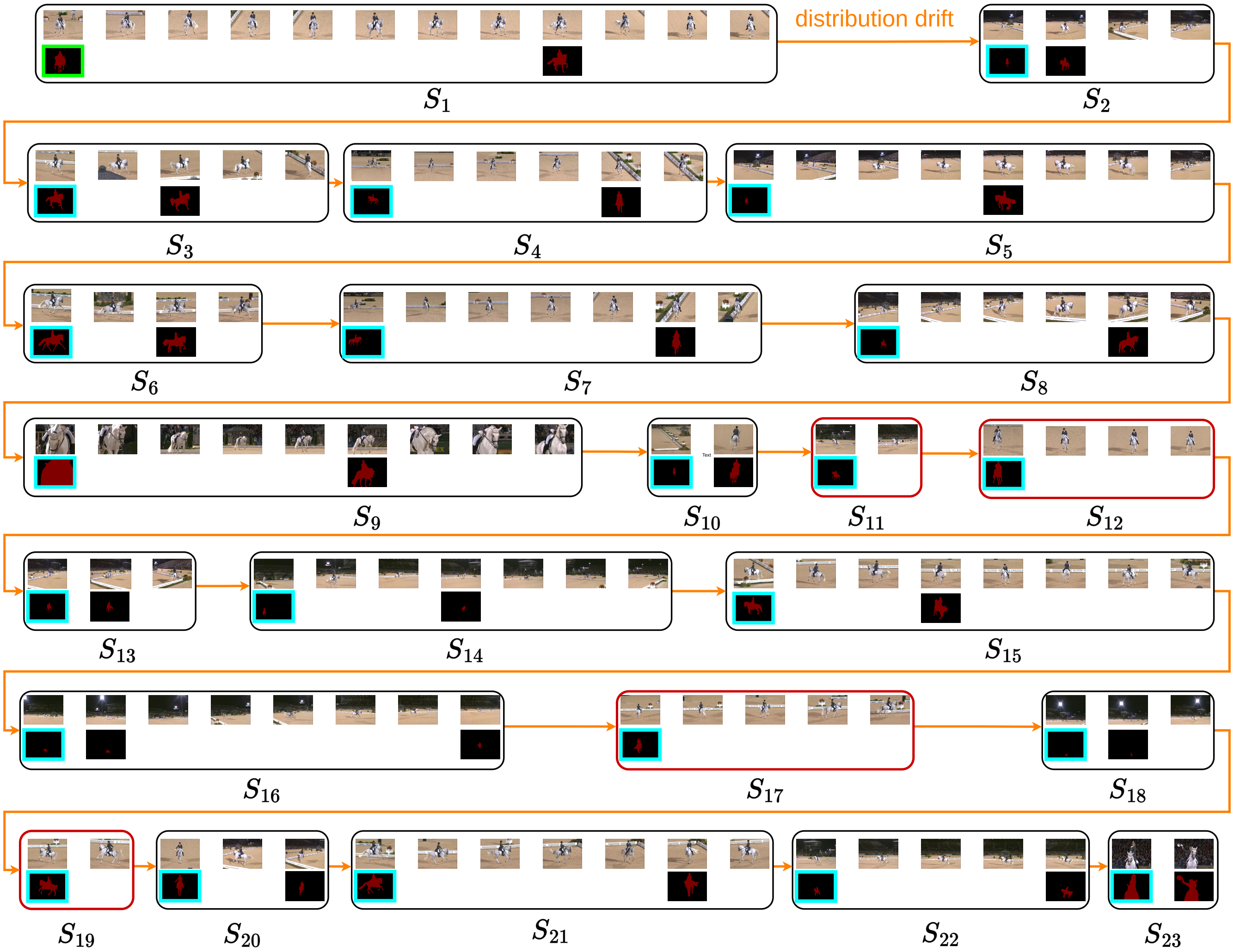

Figure 3. A subset of frames from "dressage" video of the Long Videos dataset [24]. The video consists of 23 sub-chunks that are separated from each other by significant distributional drifts or discontinuities. The lower (sparse) row, in each set, show the annotated frames. The annotations provided by [24] are shown without a border, whereas the annotated masks added via this paper, and made available via the CLVOS23 dataset, are shown with blue borders. The four sub-chunks that are missing from the Long Videos dataset are encircled in red.

6. Experimental Result

A fixed setup is used for the evaluated methods, with maximum memory sizes of $N = 32$ for LWL and LWL-RCL as suggested in LWL's original publication. For all experiments, the target model C is updated for three epochs on the memory $\mathcal{M}$ in each updating step to have a fair comparison with the baseline. The target model is updated every time the memory is updated, following the proposed setup in [10].

The memory $\mathcal{M}^0$ is initialized by the given ground truth frame F_g. In all of the experiments, as suggested in the semi-supervised online VOS baseline (LWL), the information extracted from F_g is preserved and is used throughout the evaluation of other frames in the video sequence. In the proposed method, the same concept is followed where in the proposed regularisation-based LWL, the importance parameters Φ^0 and the parameters Θ^0 related to the training of the target model C on X_g and Y_g are kept in $\mathcal{M}_R$.

In the RCL method, λ is set to 5 and the maximum size of $\mathcal{M}_R$ is set to 20. We validate these hyper-parameter using cross validation. In LWL, the target model C is a small one layer convolutional neural network. Additionally, the same pretrained decoder D and encoder models are used for all experiments of LWL. To measure the effectiveness of the proposed method, consistent with the standard DAVIS protocol [35] the mean Jaccard $\mathcal{J}$ index, mean boundary $\mathcal{F}$ scores, and the average of $\mathcal{J}\&\mathcal{F}$ are reported for all evalu-

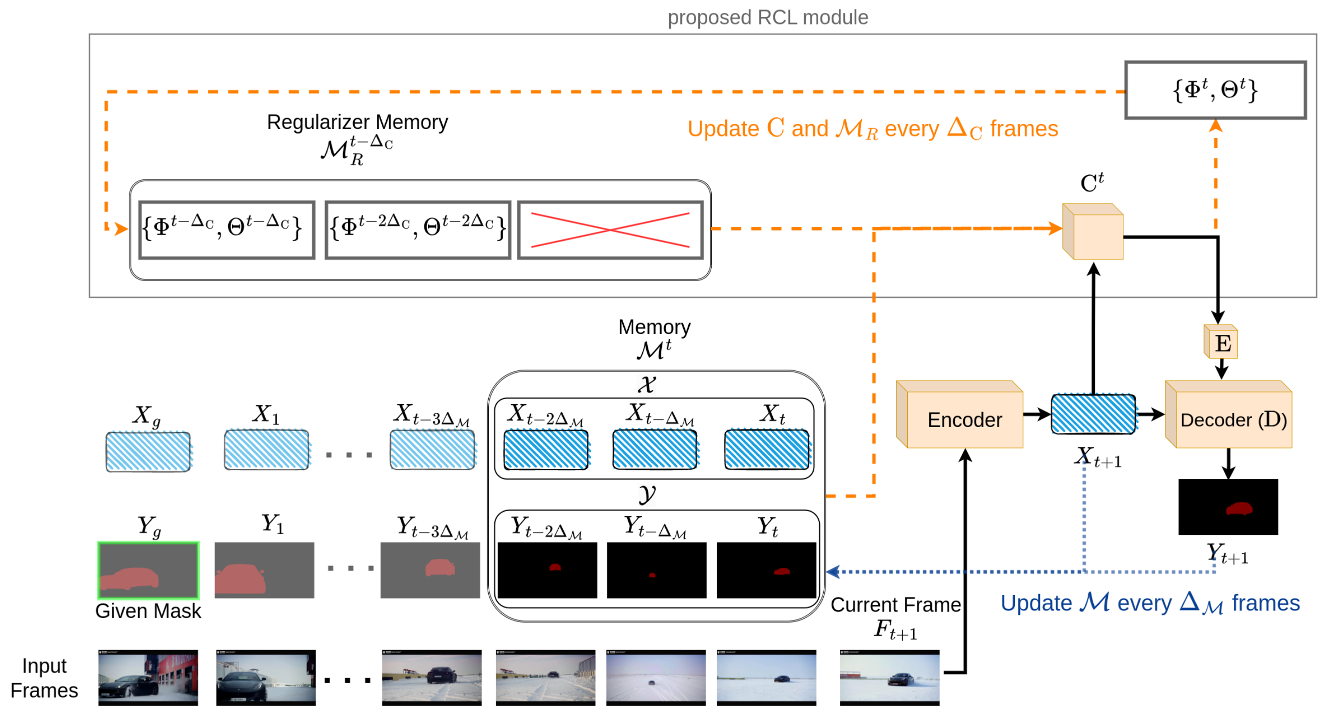
proposed RCL module
$\{\Phi^t, \Theta^t\}$
Regularizer Memory
$\mathcal{M}_R^{t-\Delta_C}$
Update C and $\mathcal{M}_R$ every Δ_C frames
$\{\Phi^{t-\Delta_C}, \Theta^{t-\Delta_C}\}$
$\{\Phi^{t-2\Delta_C}, \Theta^{t-2\Delta_C}\}$
C^t
E
Memory
$\mathcal{M}^t$
$\mathcal{X}$
X_g
X_1
$X_{t-3\Delta_\mathcal{M}}$
$X_{t-2\Delta_\mathcal{M}}$
$X_{t-\Delta_\mathcal{M}}$
X_t
Encoder
Decoder (D)
X_{t+1}
$\mathcal{Y}$
Y_g
Y_1
$Y_{t-3\Delta_\mathcal{M}}$
$Y_{t-2\Delta_\mathcal{M}}$
$Y_{t-\Delta_\mathcal{M}}$
Y_t
Y_{t+1}
Given Mask
Current Frame
F_{t+1}
Update $\mathcal{M}$ every $\Delta_\mathcal{M}$ frames
Input
Frames

Figure 4. The proposed online VOS framework, with the proposed RCL approach: At time t, the process of updating $C^{t-\Delta_C}$ on $\mathcal{M}^t$ is regularized by all pairs of the target model's parameters and their associated importance $\{\Phi, \Theta\}$ in the regularizer memory $\mathcal{M}_R^{t-\Delta_C}$ as shown in Eq. (3). After updating $C^{t-\Delta_C}$ to C^t, $\mathcal{M}_R^{t-\Delta_C}$ is updated using $\{\Phi^t, \Theta^t\}$ calculated from C^t.

ated methods. The speed of each method is reported on the DAVIS16 dataset [37] in units of Frames Per Second (FPS). All experiments were performed using one NVIDIA V100 GPU.

The effectiveness of the proposed regularization-based continual learning method (RCL) is evaluated by augmenting an online VOS framework (LWL); however, the proposed method can be extended to any online VOS method having a periodically-updated network model, as in Figure 1.

Table 2 shows the results of the selected baseline (LWL) and the augmented baseline with the proposed regularization-based method (RCL) on the Long Video dataset [24], and the proposed CLVOS23 dataset. Here, six experiments with six different memory and target model update step sizes $\Delta_C \in \{1, 2, 4, 6, 8, 10\}$ are conducted ($\Delta_\mathcal{M} = \Delta_C$), where, the memory $\mathcal{M}^t$ is updated after each target model $C^{t-\Delta_C}$ update to C^t. For reference, the means and standard deviations of six runs of two competing methods (LWL and LWL-RCL) are reported in Table 2. As it is represented in Table 2, CLVOS23 is a more difficult VOS dataset in comparison to the Long Videos dataset, since LWL has lower performance on CLVOS23. Additionally, the proposed RCL improves LWL on CLVOS23 more than the Long Videos dataset, which shows CLVOS23 is a more appropriate dataset for evaluating online, continual learning-based contributions.

Furthermore, looking at the standard deviations reported in Table 2, the proposed regularization-based method decreases the standard deviation of reported results with different memory and target model step sizes $\Delta_C \in \{1, 2, 4, 6, 8, 10\}$. This indicates that the proposed method is more robust against selecting different frame rates for updating the target model C.

Table 3 shows the results of the selected baseline on two short VOS datasets (DAVIS16 and DAVIS17). The results show that the proposed RCL method does not have any negative effects on the accuracy of the baseline method (LWL); however, it affects the speed of the baseline since it needs to recalculate the regularization term in Eq. (3) in every epoch of the updating step.

It is worth mentioning that we use the suggested hyper-parameters in the original paper of LWL [5]; nevertheless, the used hyper-parameters are not necessarily the best parameters for LWL on long video datasets, and it is possible to improve the performance of the baseline method on the evaluated dataset by only making some small changes to LWL. The objective of this article is to provide a contin-

Method	Long Videos [24]			CLVOS23		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
LWL [5]	78.0±4.3	81.6±4.2	79.8±4.2	68.1±2.2	71.9±2.4	70.0±2.3
LWL-RCL (ours)	79.8±3.0	82.7±3.2	81.3±3.1	70.4 ±1.9	74.33±2.0	72.4±2.0

Table 2. Performance analysis of the evaluated methods against the validation set of the Long Videos and proposed CLVOS23 datasets.

Method	DAVIS17			DAVIS16			FPS
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	
LWL [5]	77.1	82.9	80.0	87.3	88.5	87.9	18.15
LWL-RCL (ours)	77.1	82.9	80.0	87.3	88.5	87.9	14.47

Table 3. Performance analysis of the evaluated methods against validation sets of the DAVIS16 and DAVIS17 datasets.

ual learning-based VOS dataset and a method that improves any online VOS approaches that struggle with forgetting on long video sequences with abrupt changes in the target object's appearance.

7. Conclusion

In this article, we presented a dataset called CLVOS23 to examine the capability of semi-supervised VOS approaches to deal with the forgetting of past frames' learning, and we frame this problem as a continual learning challenge. To help online VOS methods get around memory limitations without sacrificing accuracy, we also proposed adding a regularization-based module to them. The proposed modules can be added to any existing online VOS framework that is already in place to make it more efficient and resistant to distribution drifts that can happen during long video clips, while keeping or even improving performance accuracy. The changes we made to the standard procedure for online VOS made it more accurate on long videos, according to our results. Furthermore, on the short video datasets (DAVIS16, DAVIS17) where the object's appearance does not suddenly change, the proposed methods do not outperform the baselines.

Acknowledgments

We appreciate the generous support provided by Microsoft Office Media Group and NSERC Alliance for this research project.

References

[1] Rahaf Aljundi. *Continual Learning in Neural Networks*. PhD thesis, KU Leuven, 2019. 2

[2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018. 2

[3] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019. 1

[4] Craig Atkinson, Brendan McCane, Lech Szymanski, and Anthony Robins. Pseudo-rehearsal: Achieving deep reinforcement learning without catastrophic forgetting. *Neurocomputing*, 428:291–307, 2021. 2

[5] Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *European Conference on Computer Vision*, pages 777–794. Springer, 2020. 1, 2, 3, 5, 7, 8

[6] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017. 2

[7] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018. 2

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 2

[9] Pei-Hung Chen, Wei Wei, Cho-Jui Hsieh, and Bo Dai. Overcoming catastrophic forgetting by bayesian generative regularization. In *International Conference on Machine Learning*, pages 1760–1770. PMLR, 2021. 2

[10] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 1, 2, 4, 6

[11] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2

[12] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021. 2

[13] 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. 1

[14] Zeki Erdem, Robi Polikar, Fikret Gurgen, and Nejat Yumusak. Ensemble of svms for incremental learning. In *International Workshop on Multiple Classifier Systems*, pages 246–256. Springer, 2005. 2

[15] Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018. 2

[16] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014. 2

[17] Yen-Chang Hsu, Yen-Cheng Liu, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*, 2018. 2

[18] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4144–4154, 2021. 2

[19] Amlan Kar, Seung Wook Kim, Marko Boben, Jun Gao, Tianxing Li, Huan Ling, Zian Wang, and Sanja Fidler. Toronto annotation suite. https://aidemos.cs.toronto.edu/toras, 2021. 5

[20] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2, 5

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 1

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998. 1

[23] Sang-Woo Lee, Chung-Yeon Lee, Dong-Hyun Kwak, Jiwon Kim, Jeonghee Kim, and Byoung-Tak Zhang. Dual-memory deep learning architectures for lifelong learning of everyday human behaviors. In *IJCAI*, pages 1669–1675, 2016. 2

[24] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. *Advances in Neural Information Processing Systems*, 33:3430–3441, 2020. 1, 4, 5, 6, 7, 8

[25] Fanchao Lin, Hongtao Xie, Yan Li, and Yongdong Zhang. Query-memory re-aggregation for weakly-supervised video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2038–2046, 2021. 2

[26] Yu Liu, Lingqiao Liu, Haokui Zhang, Hamid Rezatofighi, Qingsen Yan, and Ian Reid. Meta learning with differentiable closed-form solver for fast video object segmentation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8439–8446. IEEE, 2020. 2

[27] Yong Liu, Ran Yu, Jiahao Wang, Xinyuan Zhao, Yitong Wang, Yansong Tang, and Yujiu Yang. Global spectral filter memory network for video object segmentation. In *European Conference on Computer Vision*, pages 648–665. Springer, 2022. 2

[28] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26, 2017. 1

[29] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9670–9679, 2021. 1, 2

[30] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9650–9659. IEEE, 2021. 2

[31] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 2

[32] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017. 2

[33] 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*, pages 9226–9235, 2019. 2

[34] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2663–2672, 2017. 2

[35] 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, 4, 6

[36] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990. 2

[37] Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Learning fast and robust target models for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7406–7415, 2020. 1, 2, 3, 5, 7

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[39] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 2

[40] 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*, pages 12889–12898, 2021. 2

[41] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In

Advances in Neural Information Processing Systems, pages 2990–2999, 2017. 2

[42] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017. 2

[43] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 374–382, 2019. 2

[44] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1286–1295, 2021. 2

[45] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 1

[46] 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. 2

[47] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021. 2

[48] Rui Yao, Guosheng Lin, Shixiong Xia, Jiaqi Zhao, and Yong Zhou. Video object segmentation and tracking: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4):1–47, 2020. 1

[49] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13208–13217, 2020. 2

[50] Zhishan Zhou, Lejian Ren, Pengfei Xiong, Yifei Ji, Peisen Wang, Haoqiang Fan, and Si Liu. Enhanced memory network for video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2