Breaking the "Object" in Video Object Segmentation

Pavel Tokmakov Jie Li Adrien Gaidon
Toyota Research Institute
first.last@tri.global

Abstract

The appearance of an object can be fleeting when it transforms. As eggs are broken or paper is torn, their color, shape and texture can change dramatically, preserving virtually nothing of the original except for the identity itself. Yet, this important phenomenon is largely absent from existing video object segmentation (VOS) benchmarks. In this work, we close the gap by collecting a new dataset for Video Object Segmentation under Transformations (VOST). It consists of more than 700 high-resolution videos, captured in diverse environments, which are 20 seconds long on average and densely labeled with instance masks. We adopt a careful, multi-step approach to ensure that these videos focus on complex object transformations, capturing their full temporal extent. We then extensively evaluate state-of-the-art VOS methods and make a number of important discoveries. In particular, we show that existing methods struggle when applied to this novel task and that their main limitation lies in over-reliance on static appearance cues. This motivates us to propose a few modifications for the top-performing baseline that improve its capabilities by better modeling spatio-temporal information. More broadly, our work highlights the need for further research on learning more robust video object representations.

Nothing is lost or created, all things are merely transformed.

Antoine Lavoisier

1. Introduction

Spatio-temporal cues are central in segmenting and tracking objects in humans, with static appearance playing only a supporting role [23, 27, 43]. In the most extreme scenarios, we can even localize and track objects defined by coherent motion alone, with no unique appearance whatsoever [20]. Among other benefits, this appearance-last approach increases robustness to sensory noise and enables object permanence reasoning [41]. By contrast, modern computer vision models for video object segmentation [3, 11, 44, 64] operate in an appearance-first paradigm. Indeed, the most successful approaches effectively store patches with associated instance labels and retrieve the closest patches to segment the target frame [11, 38, 44, 64].

Figure 1. Video frames from the DAVIS'17 dataset [42] (above), and our proposed VOST (below). While existing VOS datasets feature many challenges, such as deformations and pose change, the overall appearance of objects varies little. Our work focuses on object transformations, where appearance is no longer a reliable cue and more advanced spatio-temporal modeling is required.

What are the reasons for this stark disparity? While some are algorithmic (e.g., object recognition models being first developed for static images), a key reason lies in the datasets we use. See for instance the "Breakdance" sequence from the validation set of DAVIS'17 [42] in Figure 1: while the dancer's body experiences significant deformations and pose changes, the overall appearance of the person remains constant, making it an extremely strong cue.

However, this example – representative of many VOS datasets – covers only a narrow slice of the life of an object. In addition to translations, rotations, and minor deformations, objects can transform. Bananas can be peeled, paper can be cut, clay can be molded into bricks, etc. These transformations can dramatically change the color, texture, and shape of an object, preserving virtually nothing of the orig-

Figure 2. Representative samples from VOST with annotations at three different time steps (see video for full results). Colours indicate instance ids, with grey representing ignored regions. VOST captures a wide variety of transformations in diverse environments and provides pixel-perfect labels even for the most challenging sequences.

inal except for the identity itself (see Figure 1, bottom and Figure 2). As we show in this paper, tracking object identity through these changes is relatively easy for humans (e.g. labelers), but very challenging for VOS models. In this work, we set out to fill this gap and study the problem of segmenting objects as they undergo complex transformations.

We begin by collecting a dataset that focuses on these scenarios in Section 3. We capitalize on the recent large-scale, ego-centric video collections [13, 21], which contain thousands of examples of human-object interactions with activity labels. We carefully filter these clips to only include major object transformations using a combination of linguistic cues (change of state verbs [19, 29]) and manual inspection. The resulting dataset, which we call VOST (Video Object Segmentation under Transformations), contains 713 clips, covering 51 transformations over 155 object categories with an average video length of 21.2 seconds. We then densely label these videos with more than 175,000 masks, using an unambiguous principle inspired by spatio-temporal continuity: if a region is marked as an object in the first frame of a video, all the parts that originate from it maintain the same identity (see Figure 2).

Equipped with this unique dataset, we analyze state-of-the-art VOS algorithms in Section 4. We strive to include a representative set of baselines that illustrates the majority of the types of approaches to the problem in the literature, including classical, first frame matching methods [61], local mask-propagation objectives [26], alternative, object-level architectures [3], and the mainstream memory-based models [11, 63–65]. Firstly, we observe that existing methods are indeed ill-equipped for segmenting objects through complex transformations, as illustrated by the large (2.3-12.5 times) gap in performance between VOST and DAVIS'17 (see Table 2). A closer analysis of the results reveals the following discoveries: (1) performance of the methods is inversely proportional to their reliance on static appearance cues; (2) progress on VOST can be achieved by improving the spatio-temporal modeling capacity of existing architectures; (3) the problem is not easily solvable by training existing methods on more data.

We conclude in Section 5 by summarizing the main challenges associated with modeling object transformations. We hope that this work will motivate further exploration into more robust video object representations. Our dataset, source code, and models are available at vostdataset.org.

2. Related Work

In this work, we study the problem of *video object segmentation* under transformations and analyze existing *VOS methods* under this novel task. Our efforts are motivated by observations about *object perception in humans*. Below, we review the most relevant works on each of these topics.

Video object segmentation is defined as the problem of pixel-accurate separation of foreground objects from the background in videos [30, 40, 51]. What constitutes foreground is either defined by independent motion [7, 40] or using a mask manually provided in the first frame of a video [30, 40, 51], the latter setting known as semi-supervised VOS. The earliest datasets lacked in scale and consistency [7, 30, 51]. The release of the DAVIS benchmark [40] was a significant step for the community as it provided 50 high-resolution sequences featuring a variety of challenges. While DAVIS caused a flurry of novel VOS methods [8, 39, 50, 54], it treated VOS as a binary foreground/background separation problem.

In contrast DAVIS'17 [42] not only extended the dataset to 150 videos, but, most importantly, introduced instance labels. In this, now de-facto standard, setting, an algorithm is provided with several object masks in the first frame and has to output pixel-perfect masks for these objects for the remainder of the video, together with their identity. While DAVIS focused on the data quality, it lacked in quantity, forcing most methods to resort to pre-training on static images [28, 39], or synthetic videos [50]. This issue was ad-

Dataset	Videos	Frames	Avg len. (s)	Masks/frame	Ann fps.	Granularity	Focus
DAVIS'16 [40]	50	3,455	3.0s	1.0	24	Binary	Data quality
DAVIS'17 [42]	150	10,700	3.0s	3.0	24	Instance	Instance labels
YTVOS [59]	3,500	120,400	4.6s	1.6	6	Instance	Dataset size
UVO* [55]	10,337	30,500	3.0s	8.8	1	Instance	Object vocabulary
VISOR [14]	7836	50,700	12.0s	5.3	0.5	Semantic	Object manipulation†
VOST (Ours)	713	75,547	21.2s	2.3	5	Instance	Object transformation

Table 1. Statistics of major video object segmentation datasets (*: train/val public annotations; †: including a small fraction of object transformation annotations). Unlike all existing VOS benchmarks, VOST focuses on the specific challenge of modeling complex object transformations. This motivates our design decisions to densely label relatively long videos with instance masks.

dressed by the large-scale YouTube-VOS benchmark [59], which features 3,252 videos over 78 categories.

Very recently, to further scale the datasets while keeping the annotation costs manageable, several works proposed to label videos at a very low fps (1 in [55] and ∼0.5 in [14]) and interpolate ground truth labels to obtain dense annotations. The estimated labels are then automatically filtered to keep only the confident interpolations. While this approach was shown to work well in many cases, in the appendix we demonstrate that it fails precisely in the most challenging scenarios which we are interested in.

Notably, none of these datasets features a significant amount of object transformations. Thus, our effort is complementary to existing work. We compare VOST to major VOS benchmarks in Table 1, illustrating our key design decisions. In particular, we label relatively long videos to capture the full extent of each transformation, and provide temporally dense instance-level labels, as interpolation fails when objects transform.

VOS methods can be categorized in many possible ways. Here we focus on the semi-supervised setting and trace the history of the field to identify main trends. Early, pre-deep learning methods propagate the first frame labels over a spatio-temporal graph structure by optimizing an energy function [4, 18, 22], but struggle with generalization due to their heuristic-based nature.

First deep-learning solutions had to deal with the lack of video data for training and hence modeled video segmentation as an image-level problem [8, 28, 57]. In particular, these works proposed to pre-train a CNN for binary object segmentation on COCO [34] and then fine-tuned the model for a few iterations on the first frame of a test video. While this approach outperformed heuristic-based methods, it is computationally expensive and not robust to appearance change. These issues were separately addressed in [10, 25, 61], which replace expensive fine-tuning with cheap patch-level matching, and in [31, 35, 39, 54] which introduce online adaptation mechanisms.

More recently, memory-based models have become the mainstream approach for semi-supervised video object segmentation [11, 37, 38, 44, 53, 63–65]. The earliest methods in this category [37, 53, 63] extend the first-frame matching mechanism of [10, 25, 61] by additionally matching with the previous frame. This architecture can be seen as a memory module with capacity 2, providing an efficient mechanism for adapting to appearance changes. More advanced versions of the architecture include increasing the memory capacity by storing several previous frames [38, 44], using transformers [15, 52] for retrieving object labels from memory [16, 64], introducing memory compression to support longer sequences [11, 33], and improving the efficiency of the memory read operation [12, 45, 58].

Alternative approaches to VOS include supervised [9, 24, 39] and, more recently, unsupervised [26, 56] mask propagation methods that do not maintain an appearance model of the target. These methods are very efficient, but cannot handle occlusions and suffer from drift in longer sequences. A few works [3, 32, 66] propose to perform appearance matching on the object, not on the patch level, but their accuracy remains low. Finally, coherent motion is a key signals for object perception in humans, but it was mostly studied in unsupervised VOS [50, 60, 62].

In this work, we evaluate a representative set of semi-supervised VOS methods on the task of segmenting objects as they undergo complex transformations. Our experiments illustrate limitations of the appearance-first paradigm, motivating the exploration of spatio-temporal architectures.

Object perception in humans is driven by spatio-temporal cohesion. At the early development stages, infants use the notions of boundedness and cohesion in space-time, not static, gestalt cues like shape or texture to group surfaces into objects [47–49]. In adults, the object files theory [27] postulates that our visual system *individuates* each object by grouping visual primitives based on spatio-temporal factors. Most importantly, object's individuation precedes its appearance identification, as shown in [23, 27, 43]. That is, humans can perceive something as the same 'thing' while its appearance remains in flux and might dramatically change over time. In the most extreme cases, individuation can function in the absence of any unique object appearance, as shown by Gao and Sholl [20].

Very recently, Peters and Kriegeskorte [41] summarized

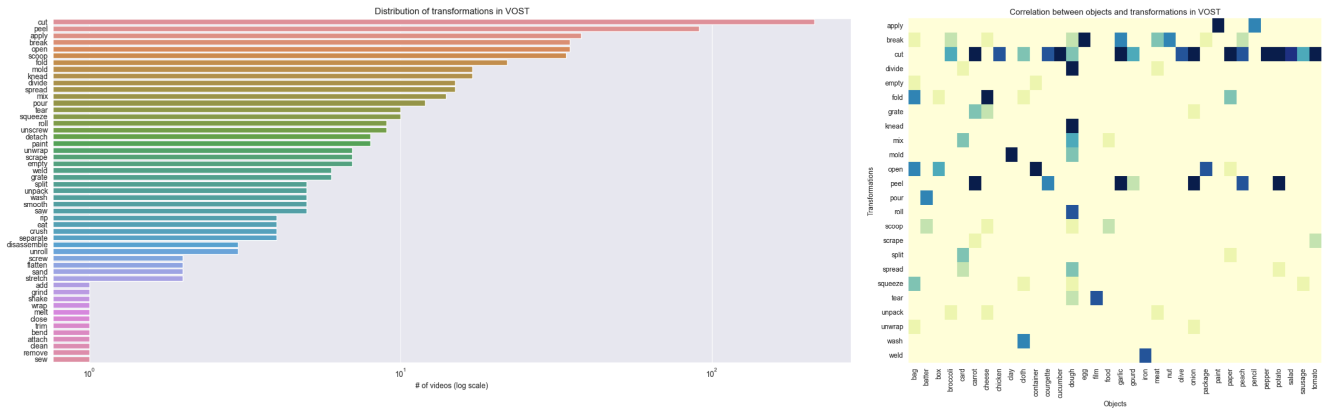

Figure 3. Statistics of VOST: distribution of transformations on the left, and co-occurrence statistics between the most common transformations and object categories on the right. While there is some bias towards common activities, like cutting, the tail of the distribution is sufficiently heavy. Moreover, cutting has a broad semantic meaning, resulting in diverse transformations. Best viewed with zoom.

the differences between object representations in the brain and neural networks, including the dichotomy between spatio-temporal and appearance cues. They then argue that the best way to bridge the differences between these two types of representations is by introducing novel machine vision tasks that require more complex spatio-temporal reasoning. In this work, we make a step in this direction by extending the setting of video object segmentation to support object transformations.

3. Dataset Design and Collection

In this section, we discuss our approach to collecting VOST. The key steps include selecting representative videos, annotating them with instance masks, and defining an evaluation protocol.

3.1. Video selection

We choose to source our videos from the recent large-scale, egocentric action recognition datasets, which provide temporal annotations for a large vocabulary of activities. In particular, we use EPIC-KITCHENS [13] and Ego4D [21], where the former captures activities in kitchens, such as cooking or cleaning, and the later provides a much larger diversity of scenarios, including outdoor ones. It is worth noting that the egocentric focus of VOST is merely an artifact of the datasets that were used to source the videos. The nature of the problem itself is independent of the camera viewpoint and we expect that approaches developed on VOST will generalize to third-person videos.

While these datasets feature tens of thousands of clips, the vast majority of the actions (e.g., 'take' or 'look') do not result in object transformations. To automatically filter out such irrelevant clips, we capitalize on the notion of change of state verbs from the language theory [19, 29]. That is, rather than manually filtering the videos themselves, we first filter the action labels. This dramatically reduces the total number of clips we have to consider to 10,706 (3,824 from EPIC-KITCHENS and 6,882 from Ego4D).

Although all the clips selected above feature an object state change, not all result in a significant appearance change. For example, folding a towel in half, or shaking a paintbrush does nearly nothing to their overall appearance. To focus on the more challenging scenarios, we manually review each video and label its complexity on a scale from 1 to 5, where 1 corresponds to no visible object transformation and 5 to a major change of appearance, shape and texture (see appendix for details). In addition, at this stage we merge clips representing several steps of the same transformation (e.g. consecutive cuts of an onion). After collecting these labels we find that the majority of videos in the wild are not challenging, however, we are still left with 986 clips in the 4-5 range, capturing the entire temporal extent of these complex transformations.

Finally, we further filter the clip based on two criteria. Firstly, some videos are nearly impossible to label accurately with dense instance masks (e.g., due to excessive motion blur), so we skip them. Secondly, there are a few large clusters of near duplicates (e.g., there are 116 clips of molding clay into bricks that are performed by the same actor in the same environment), so we sub-sample those to reduce bias. The resulting dataset contains 713 videos covering 51 transformations over 155 object categories. Note that, in accordance with the standard VOS protocols [42, 59], semantic labels are only used for data collection and are not provided as input to the algorithms.

The distribution over transformations and co-occurrence statistics between transformations and objects are shown in Figure 3. Firstly we observe that, although there is some bias towards more common actions, such as cutting, the long tail of interactions is sufficiently heavy. Moreover, as evident from the correlation statistics on the right side of

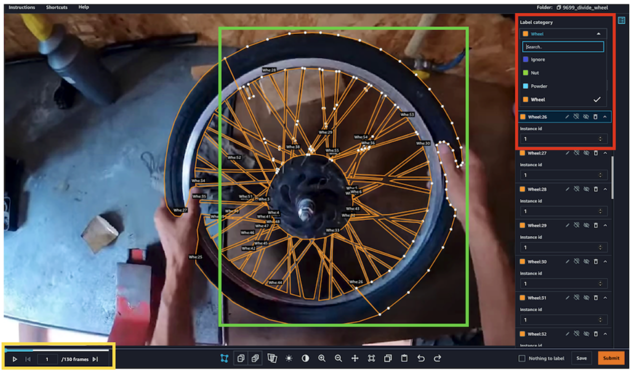

Figure 4. Interface of our annotation tool. Objects are annotated with polygons (shown in green), and additional "Category" and "Instance id" labels (red). Annotations are automatically propagated to the next frame and then manually adjusted (yellow).

the figure, cutting has an extremely broad semantic meaning and can be applied to almost any object, resulting in very different transformations (see cutting corn and paper in Figure 2). Overall, there is substantial entropy in the correlation statistics illustrating the diversity of our dataset.

3.2. Annotation collection

To label the videos selected above, we begin by adjusting the temporal boundaries of each clip to tightly enclose the entire duration of the transformation, with the exception of extremely long sequences (a minute or longer). To balance the cost and temporal density of the annotations we choose to label videos at 5 fps.

A key question is how to annotate objects as they split into parts (e.g. due to cutting or breaking). To avoid ambiguity, we adopt the most straightforward and general principle: if a region is marked as an object in the first frame of a video, all the parts that originate from it maintain the same identity. For example, the yolks from the broken eggs in Figure 2 maintain the identity of the object they originated from. This approach also ensures that there is an unambiguous signal in the data (spatio-temporal continuity) that algorithms can use to achieve generalization.

There are, however, examples in which it is impossible to provide an accurate instance mask for a region. In the second row of Figure 2 we show two such cases. In the first one, a piece of clay is experiencing fast motion, making establishing a clear boundary impossible. In the second example, the egg whites from several eggs are mixed together, making it impossible to separate them from each other. Rather than skipping such videos, we choose to label the ambiguous regions with tight "Ignore" segments (shown in gray in the figure), which are not used at either training or evaluation time. This flexible approach allows us to consistently annotate even the most challenging videos.

Given the complexity of the task, we hired a fixed team of 20 professional annotators for the entire duration of the project. They received detailed instructions on the task and edge cases which we detail in the appendix. The annotators were first trained for 4 weeks to ensure consistent behavior. Each video was labeled by one annotator using Amazon SageMaker GroundTruth tool for polygon labeling shown in Figure 4. For videos featuring multiple objects and an additional "Instance id" label was provided. The videos were then reviewed by a small, held-out group of skilled annotators and returned to the original worker for correction. This process was repeated until no more issues could be identified. On average, 3.9 annotation-review cycles were performed for each video to ensure the highest label quality.

Overall 175,913 masks were collected, with an average track duration of 21.3 seconds. We report additional statistics of the dataset in the appendix.

3.3. Splits and metrics

VOST is split into 572 train, 70 validation, and 71 test videos. We have released the labels for train and validation sets, but the test set is held out and only accessible via an evaluation server to prevent over-fitting. Furthermore, we ensure that all three sets are well separated by enforcing that each kitchen from [13] and each subject from [21] appears in only one of the train, validation or test sets.

For evaluation, traditionally, video object segmentation datasets use a combination of region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$ [42, 59]. The former is the standard intersection-over-union [17] between the predicted M and ground truth masks G, which captures the fraction of pixels that are correctly labeled. Contour accuracy, on the other hand, measures how accurate the boundaries of the predicted masks are [36]. Both quantities are computed separately for each instance in each frame and then averaged over frames in a video and over instances.

We propose two modifications to the standard metrics to better reflect our problem setting. Firstly, we note that contours are often not well defined for the kind of masks we are dealing with: some objects are semi-transparent, and the amount of motion blur is significant. Thus, we do not measure contour accuracy in our experiments. Secondly, recall that region similarity $\mathcal{J}$ for every object o_i is averaged over all video frames:

$$\mathcal{J}(o_i, F) = \frac{1}{|F|} \sum_{f \in F} \mathcal{J}(M^f_{o_i}, G^f_{o_i}), \tag{1}$$

where F is the set of frames and $M^f_{o_i}, G^f_{o_i}$ are the predicted and ground truth masks for object o_i in frame f respectively. Hence, every frame has an equal influence on the overall score. This is adequate for the standard VOS setting, but we are interested not in how well a method can segment an object overall, but in how robust it is to transformations. To reflect this fact, we separately measure the region similarity

	VOST val		VOST test		DAVIS'17 val	
	$\mathcal{J}_{tr}$	$\mathcal{J}$	$\mathcal{J}_{tr}$	$\mathcal{J}$	$\mathcal{J}_{tr}$	$\mathcal{J}$
OSMN Match [61]	7.0	8.7	8.5	10.2	41.3	49.6
OSMN Tune [61]	17.6	23.0	20.1	26.1	57.2	68.3
CRW [26]	13.9	23.7	20.8	28.0	53.6	64.4
CFBI [63]	32.0	45.0	32.1	43.9	75.0	79.3
CFBI+ [65]	32.6	46.0	31.6	46.7	76.3	80.1
AOT [64]	**36.4**	**48.7**	**37.1**	**49.9**	80.4	82.3
XMem [11]	33.8	44.1	32.0	44.0	**81.1**	**82.9**
HODOR Img [3]	13.9	24.2	22.1	29.0	70.2	74.7
HODOR Vid [3]	25.4	37.1	27.6	42.0	74.0	77.4

Table 2. Benchmarking existing methods on VOST. We report results on both validation and test sets of our dataset, using IoU after transformation $\mathcal{J}_{tr}$ as well as the overall IoU $\mathcal{J}$. We include DAVIS'17 val scores for reference. Performance of all methods is 2.2-5.9 times lower in terms of $\mathcal{J}_{tr}$ on VOST compared to DAVIS, emphasizing the complexity of the problem.

after the transformation has been mostly completed: $\mathcal{J}_{tr} = \mathcal{J}(o_i, \hat{F})$, where $\hat{F}$ represents the last 25% of the frames in a sequence. We report both $\mathcal{J}$ and $\mathcal{J}_{tr}$ in our experiments, but use the latter as the main metric.

4. Analysis of the State-of-the-art Methods

We now use VOST to analyze how well can existing VOS methods handle object transformations. All the models are initialized from their best DAVIS'17 checkpoint (usually pre-trained on a large-scale image and/or video collection) and fine-tuned on the training set of VOST, unless stated otherwise. We use the original implementations, only adapting the loss to correctly handle "Ignore" labels and tuning the number of training iterations on the validation set. More details are provided in the appendix.

4.1. Methods

We evaluate a total of nine video segmentation algorithms and their variants, which are selected to cover the main trends in the field over recent years. In addition, the methods' performance on existing benchmarks and public availability of the code were taken into account.

We include OSMN [61] as a representative approach for early deep-learning methods that either fine-tune a CNN on the first frame (denoted as OSMN Tune) or employ a more efficient matching mechanism (OSMN Match). As a complementary approach, we evaluate the self-supervised CRW objective [26] for mask propagation which only uses local information between consecutive frame pairs.

In the mainstream, memory-based family of methods we evaluate CFBI [63] and its improved variant CFBI+ [65], which have been established as very strong baselines on existing benchmarks. In addition, we include the transformer-based AOT approach [64], and the very recent XMem framework [11], which specifically focuses on long videos.

Finally, we study another recent method - HODOR [3], which performs template matching on the object, not on the patch level. We include both the image-based version of this approach, which is trained on COCO (denoted as HODOR Img), as well as the video-based one (HODOR Vid).

4.2. Results

Can existing methods handle transformations? In Table 2 we start by reporting the performance of approaches described above on the validation and test sets of VOST. For reference, we also report the performance of these methods on the validation set of DAVIS'17 on the right.

Firstly, we observe that the appearance matching baseline (OSMN Match in the table) fails dramatically. This is to be expected as virtually all videos in our dataset feature major appearance changes. Expensive test time fine-tuning on the first frame of a video (OSMN Tune) improves the performance of this baseline, but the validation set score remains 3.3 times lower than on DAVIS. Local mask propagation used by CRW is more robust to appearance change, but cannot handle occlusions, which are plentiful in first-person videos, and hence also struggles on VOST.

Next, we see that the more advanced, memory-based methods (rows 4 to 7 in the table) are indeed more capable due to their efficient mechanism for updating the appearance model of the target. That said, performance remains low, with the gap between $\mathcal{J}_{tr}$ and $\mathcal{J}$ on VOST being especially large. On DAVIS, on the other hand, the gap is almost completely eliminated by the most recent AOT and XMem baselines. These results demonstrate that, while memory-based methods are capable of segmenting objects through minor appearance changes caused by translations and deformations, they fail under more challenging transformations.

Another notable observation is that the image-based HODOR baseline (HODOR Img in the table), which is only trained on COCO, shows a major loss in performance compared to DAVIS. This illustrates that static object models learned from images break when objects start to transform. Moreover, the variant of this model trained on videos also

	OSMN Tune [61]	CFBI+ [65]	AOT [64]	HODOR Vid [3]
All	17.6 (-0.0)	32.6 (-0.0)	36.4 (-0.0)	25.4 (-0.0)
LNG	12.4 (-5.2)	30.4 (-2.2)	34.7 (-1.7)	25.0 (-0.4)
MI	14.7 (-2.9)	26.4 (-6.2)	27.2 (-9.2)	20.6 (-4.8)
OCC	17.2 (-0.4)	28.1 (-4.5)	30.7 (-5.7)	17.6 (-7.8)
FM	17.0 (-0.6)	21.8 (-10.7)	23.8 (-12.5)	16.0 (-9.4)
SM	14.4 (-3.2)	23.3 (-9.2)	24.7 (-11.7)	16.6 (-8.8)

Table 3. Quantitative evaluation of failure modes of a subset of the baselines on the validation set using $\mathcal{J}_{tr}$. We analyze such factors as video length (LNG), presence of several instances (MI), occlusions (OCC), fast object motion (FM) and small objects (SM).

underperforms, indicating that object-level matching might not be the optimal approach when object shape and appearance change significantly during the video.

What makes the problem challenging? Significant change in object shape and appearance is one factor that is common to virtually all the videos in VOST. We now analyze a representative subset of the baselines more closely to identify their additional failure modes. To this end, in Table 3 we report the $\mathcal{J}_{tr}$ score on subsets of the validation set characterized by various *quantifiable* challenges, such as the length of the video, or presence of occlusions.

Firstly, by evaluating on videos that are longer than 20 seconds (indicated with LNG in the table), we observe that length alone does not present a significant challenge for most of the methods. This demonstrates that the complexity of the problem is associated with the content of our dataset (object transformations), not with the technical challenges of processing long sequences.

One unique aspect of our task is that the objects in multi-instances sequences are typically close in appearance (e.g. several eggs). Evaluating on such sequences (indicated with MI in the table) significantly reduces the performance of all the methods. It is not surprising, as appearance-first models are especially ill-suited for this scenario. Intriguingly, the object-level matching strategy of HODOR Vid as well as the expensive test time fine-tuning of OSMN Tune, although less effective overall, seems to be more robust to multi-instance segmentation.

Next, we look at two aspects that test the methods' object permanence capabilities - full occlusions (denoted as OCC in the table) and fast motion (FM), which is often associated with objects going out of frame. The latter is measured as the distance between object centers in consecutive frames normalized by the object size (see appendix for details). Interestingly, the simplest OSMN Tune baseline is the most robust to object disappearance. More advanced methods rely heavily on the objects being visible throughout the video and struggle in highly dynamic scenes.

Finally, our dataset features many small objects (denoted as SM in the table), which are equally challenging for all methods. Overall, we can conclude that reliance on appearance cues and the lack of spatio-temporal modeling capabilities (e.g. modeling object permanence) are some of the main limitations of existing approaches.

	AOT [64]	+ 15 fr.	+ R-STM	+ 10 fps.	+ m-s.
$\mathcal{J}_{tr}$	36.4	37.4	38.5	40.7	40.1
$\mathcal{J}$	48.7	49.2	49.7	51.9	52.3

Table 4. Addressing some of the limitations of AOT. We experiment with training on longer sequences, replacing short-term memory with a recurrent transformer and increasing temporal and spatial resolution.

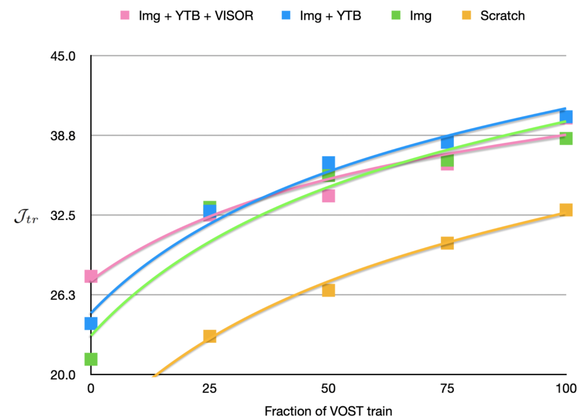

Figure 5. Evaluation of the effect of the training set size on AOT+ using $\mathcal{J}_{tr}$ on the validation set of VOST. We investigate both the effect of pre-training (on static images and videos) and the fraction of VOST train used for fine-tuning.

Are these challenges easy to address? After we have observed that VOST features many challenges that are underrepresented in existing benchmarks, it is natural to ask if we can modify the top-performing AOT baseline to address at least some of them. To this end, in Table 4 we explore several intuitive directions. Firstly, we increase the length of the training sequences from 5 to 15 frames. While this leads to some improvements, they are limited as the model is ill-equipped to capitalize on longer-term temporal cues.

Next, we increase the spatio-temporal modeling capacity of AOT by replacing the short-term memory module, which uses a transformer to match the patches in the current and previous frame, with a recurrent transformer (denoted as R-STM in the table, details are provided in the appendix). It is more similar to classical recurrent architectures like ConvGRU [5] and can aggregate a rich spatio-temporal representation of a video over time. This modification translates to stronger transformation modeling capabilities, as indicated by the improved $\mathcal{J}_{tr}$ score.

Finally, we experiment with increasing the temporal and spatial resolution of the model at test time by evaluating at 10 fps and enabling multi-scale inference (denoted as m-s. in the table). Both modifications increase the overall performance $\mathcal{J}$, but, notably, the spatial resolution has a smaller effect and even decreases $\mathcal{J}_{tr}$ somewhat. This result suggests that accurately modeling fine-grained temporal infor-

Figure 6. Qualitative results of our AOT+ baseline on sequences from validation and test sets of VOST (see video for full results). Colours represent instance ids. We can see that, while existing, appearance-first methods can handle relatively challenging transformations, they struggle in the most testing scenarios when appearance is either not enough to distinguish between objects or it changes dramatically.

mation is key for achieving progress on VOST.

We qualitatively analyze both the success and failure modes of the final variant, which we denote as AOT+, in Figure 6. Firstly, we can see that this model can perfectly handle the banana peeling sequence, illustrating its robustness to relatively challenging transformations. However, in the next sequence the limitations of appearance-first approaches start to show. AOT+ first confuses the coffeemaker with the hand due to reflection and then fails to separate the top part of the objects from the metal sink. Next, in the paper rolling sequence, the two instances are correctly segmented at first, but as soon as they are moved and folded together the track of identities is lost and the model breaks down. Finally, AOT+ fails completely in the very challenging egg-cracking example, being not able to both segment the full extents of the eggs and to distinguish between them.

Is more data all you need? We now investigate whether the challenges we saw above can be addressed by simply training a model like AOT+ on a larger dataset. To this end, we vary both the pre-training datasets and the size of the VOST training set itself and report the results in Figure 5.

Firstly, we see that although pre-training is important the static image dataset proposed in [33] is enough to provide a strong initialization and further pre-training on videos from YTVOS only brings marginal improvements. We additionally experiment with pre-training on the very recent VISOR dataset [14], which is also sourced from EPIC-KITCHENS, but features only few transformations. In-domain pre-training indeed improves zero-shot performance but does not bring noticeable benefits after fine-tuning on VOST.

Finally, we observe that, while increasing the size of the training set of VOST does have a noticeable effect on performance, the improvements quickly saturate. If we extrapolate the trend, it would require labeling at least 30,000 videos with complex object transformations for AOT+ to reach the score of 80.0 on $\mathcal{J}_{tr}$, which is not practical.

5. Discussion and Limitations

In this work, we demonstrated that segmenting objects through transformations presents novel challenges, which existing algorithms are ill-equipped to address. Our analysis provides insights into the failure modes of these methods, while further raising a number of important questions.

Ambiguity is inevitable when dealing with object transformations. When designing VOST, we have put a lot of effort to make the annotations as consistent as possible. To this end, we followed the established opinion in cognitive science literature [27, 47] that object perception is driven by universal principles, such as spatio-temporal cohesion and object permanence. We further ensured that for a few scenarios that cannot be resolved on the basis of these principles alone the "Ignore" label is used. That said, providing additional annotations, for example, in the form of semantic labels for objects parts, could further enrich the dataset.

Data plays a key role in deep learning, however, our analysis in Figure 5 demonstrates that pre-training on generic video collections does not result in significant improvements on VOST. What is needed is large amounts of data featuring object transformations. As we have shown in Section 3.1, collecting such videos requires a lot of effort. Automatic data collection using recent, self-supervised visual-language models [1, 46] is a promising way to scale the dataset and extend it to third-person videos.

Model architectures are another important dimension of the problem. The capacity of a model is what determines how effectively it can use the available data. In Table 4 we have shown that, while extending the spatio-temporal capabilities of existing approaches can help, incremental improvements do not address the most fundamental challenges. An entirely new approach to modeling objects in videos is needed, with the recent spatio-temporal transformer architectures [2, 6] being a possible candidate.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 8

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *ICCV*, 2021. 8

[3] Ali Athar, Jonathon Luiten, Alexander Hermans, Deva Ramanan, and Bastian Leibe. HODOR: High-level object descriptors for object re-segmentation in video learned from static images. In *CVPR*, 2022. 1, 2, 3, 6, 7

[4] S Avinash Ramakanth and R Venkatesh Babu. Seamseg: Video object segmentation using patch seams. In *CVPR*, 2014. 3

[5] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. In *ICLR*, 2016. 7

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 8

[7] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 2

[8] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, 2017. 2, 3

[9] Xi Chen, Zuoxin Li, Ye Yuan, Gang Yu, Jianxin Shen, and Donglian Qi. State-aware tracker for real-time video object segmentation. In *CVPR*, 2020. 3

[10] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *CVPR*, 2018. 3

[11] Ho Kei Cheng and Alexander G Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 1, 2, 3, 6

[12] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *NeurIPS*, 2021. 3

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for EPCI-KITCHENS-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. 2, 4, 5

[14] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Ely Locke Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR benchmark: Video segmentations and object relations. In *NeurIPS, Datasets and Benchmarks Track*, 2022. 3, 8

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[16] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. SSTVOS: Sparse spatiotemporal transformers for video object segmentation. In *CVPR*, 2021. 3

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5

[18] Qingnan Fan, Fan Zhong, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. JumpCut: non-successive mask transfer and interpolation for video cutout. *ACM Trans. Graph.*, 34(6):195–1, 2015. 3

[19] Charles J Fillmore. The grammar of hitting and breaking. 1967. 2, 4

[20] Tao Gao and Brian J Scholl. Are objects required for object-files? Roles of segmentation and spatiotemporal continuity in computing object persistence. *Visual Cognition*, 18(1):82–109, 2010. 1, 3

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2, 4, 5

[22] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*. IEEE, 2010. 3

[23] Andrew Hollingworth and Steven L Franconeri. Object correspondence across brief occlusion is established on the basis of both spatiotemporal and surface feature cues. *Cognition*, 113(2):150–166, 2009. 1, 3

[24] Yuan-Ting Hu, Jia-Bin Huang, and Alexander Schwing. Mask-RNN: Instance level video object segmentation. In *NeurIPS*, 2017. 3

[25] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Videomatch: Matching based video object segmentation. In *ECCV*, 2018. 3

[26] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. In *NeurIPS*, 2020. 2, 3, 6

[27] Daniel Kahneman, Anne Treisman, and Brian J Gibbs. The reviewing of object files: Object-specific integration of information. *Cognitive psychology*, 24(2):175–219, 1992. 1, 3, 8

[28] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for video object segmentation. *International Journal of Computer Vision*, 127(9):1175–1197, 2019. 2, 3

[29] Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993. 2, 4

[30] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013. 2

[31] Yuxi Li, Ning Xu, Jinlong Peng, John See, and Weiyao Lin. Delving into the cyclic mechanism in semi-supervised video object segmentation. *NeurIPS*, 2020. 3

[32] Shuxian Liang, Xu Shen, Jianqiang Huang, and Xian-Sheng Hua. Video object segmentation with dynamic memory networks and adaptive object alignment. In *ICCV*, pages 8065–8074, 2021. 3

[33] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. *NeurIPS*, 2020. 3, 8

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3

[35] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018. 3

[36] David R Martin, Charless C Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004. 5

[37] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, 2018. 3

[38] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 1, 3

[39] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, 2017. 2, 3

[40] 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 *CVPR*, 2016. 2, 3

[41] Benjamin Peters and Nikolaus Kriegeskorte. Capturing the objects of vision with neural networks. *Nature Human Behaviour*, 5(9):1127–1144, 2021. 1, 3

[42] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1, 2, 3, 4, 5

[43] Brian J Scholl. Object persistence in philosophy and psychology. *Mind & Language*, 22(5):563–591, 2007. 1, 3

[44] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *ECCV*, 2020. 1, 3

[45] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *ICCV*, 2021. 3

[46] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022. 8

[47] Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990. 3, 8

[48] Elizabeth S Spelke. Where perceiving ends and thinking begins: The apprehension of objects in infancy. In *Perceptual development in infancy*, pages 209–246. Psychology Press, 2013. 3

[49] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007. 3

[50] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *ICCV*, 2017. 2, 3

[51] David Tsai, Matthew Flagg, Atsushi Nakazawa, and James M Rehg. Motion coherent tracking using multi-label MRF optimization. *International Journal of Computer Vision*, 100(2):190–202, 2012. 2

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[53] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. FeelVOS: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019. 3

[54] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *BMVC*, 2017. 2, 3

[55] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. 3

[56] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019. 3

[57] Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. Monet: Deep motion exploitation for video object segmentation. In *CVPR*, 2018. 3

[58] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *CVPR*, 2021. 3

[59] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. YouTube-VOS: Sequence-to-sequence video object segmentation. In *ECCV*, 2018. 3, 4, 5

[60] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021. 3

[61] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K. Katsaggelos. Efficient video object segmentation via network modulation. *CVPR*, 2018. 2, 3, 6, 7

[62] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *CVPR*, 2019. 3

[63] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, 2020. 2, 3, 6

[64] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *NeurIPS*, 2021. 1, 2, 3, 6, 7

[65] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3, 6, 7

[66] Xiaohui Zeng, Renjie Liao, Li Gu, Yuwen Xiong, Sanja Fidler, and Raquel Urtasun. DMM-Net: Differentiable mask-matching network for video object segmentation. In *ICCV*, 2019. 3