Reducing the Annotation Effort for Video Object Segmentation Datasets

Paul Voigtlaender[1] Lishu Luo[2,†] Chun Yuan[2] Yong Jiang[2] Bastian Leibe[1]
[1]RWTH Aachen University [2]Tsinghua Shenzhen International Graduate School
{voigtlaender,leibe}@vision.rwth-aachen.de {lls18@mails,yuanc@sz,jiangy@sz}.tsinghua.edu.cn

Abstract

For further progress in video object segmentation (VOS), larger, more diverse, and more challenging datasets will be necessary. However, densely labeling every frame with pixel masks does not scale to large datasets. We use a deep convolutional network to automatically create pseudo-labels on a pixel level from much cheaper bounding box annotations and investigate how far such pseudo-labels can carry us for training state-of-the-art VOS approaches. A very encouraging result of our study is that adding a manually annotated mask in only a single video frame for each object is sufficient to generate pseudo-labels which can be used to train a VOS method to reach almost the same performance level as when training with fully segmented videos. We use this workflow to create pixel pseudo-labels for the training set of the challenging tracking dataset TAO, and we manually annotate a subset of the validation set. Together, we obtain the new TAO-VOS benchmark, which we make publicly available at `www.vision.rwth-aachen.de/page/taovos`*. While the performance of state-of-the-art methods on existing datasets starts to saturate, TAO-VOS remains very challenging for current algorithms and reveals their shortcomings.*

1. Introduction

Semi-supervised Video Object Segmentation (VOS) is the task of segmenting objects in every frame of a video given their ground truth masks in the first frame in which they appear. VOS is a fundamental task in computer vision with important applications including video editing, robotics, and self-driving cars. Recently, VOS has received a lot of attention in the computer vision community and the quality of results has significantly improved. In addition to the adoption of recent deep learning methods, this progress was largely driven by the introduction of the DAVIS 2017 [28] and the YouTube-VOS [37] benchmarks.

With its size of 60 videos for training and 30 videos for validation, DAVIS 2017 is very useful for measuring the performance of VOS methods, but its utility for training is limited. The introduction of the much larger YouTube-VOS dataset with 3,471 training sequences was a milestone for VOS and led to a large jump in result quality. In particular, it enabled the development of end-to-end trained VOS methods. However, the videos of YouTube-VOS are short, contain relatively low diversity with a focus mainly on persons and animals, the number of objects per video is relatively low, and the total amount of data is still much smaller than what is available for other areas of computer vision such as image classification or object detection.

Hence, we argue that in order to make significant progress in VOS, new datasets which contain more diverse data, longer sequences, and more objects per sequence will be necessary. However, manually labeling all objects in a video with segmentation masks for every frame is extremely time consuming, which limits the creation of such datasets.

Contributions. Towards the goal of creating VOS datasets with less effort, we make the following contributions.

(1) We show that densely labeling all frames in a video with segmentation masks by hand is not necessary to benefit from the additional data. Only annotating bounding boxes requires an order of magnitude less effort, and yet we show that these labels can be used to train a VOS model and achieve results which come close to what can be achieved with dense pixel-wise labeling. **(2)** We investigate a strategy of manually annotating a single segmentation mask per object (in a single frame), resulting in a huge reduction of the number of masks to be annotated. We evaluate this methodology on the YouTube-VOS dataset. Here, we only use 4% of all hand-labeled masks and are able to retain more than 97% of the performance that is achieved with full supervision, which demonstrates that this setup achieves an excellent trade-off between annotation effort and utility of the labels for training. **(3)** As a first step towards new VOS datasets, we apply this method for labeling the recent Tracking Any Object (TAO) dataset [10], which so far only had bounding box annotations, and obtain the challenging TAO-VOS benchmark, which we make publicly available. **(4)** We evaluate current state-of-the-art methods on the new TAO-VOS benchmark and show that TAO-VOS is effective

[†]Work performed at RWTH Aachen under a RWTH Aachen - Tsinghua Junior Research Fellowship.

at highlighting significant performance differences, something that is no longer possible for the already saturated DAVIS and YouTube-VOS benchmarks.

2. Related Work

In the following, we review work on learning with fewer segmentation labels and reducing annotation effort.

Weakly-Supervised Image Segmentation. In the context of semantic segmentation or instance segmentation for static images, most focus has been put on learning segmentation only from image-level classification labels [18, 3, 2].

Other methods closer to our work focus on segmentation from bounding box supervision [9, 17, 15]. However, even when using bounding boxes, the result quality still lacks behind what can be achieved with more supervision.

In contrast to the aforementioned methods, we consider videos with given bounding box annotations and we also make use of existing segmentation annotations for static image datasets, allowing us to create segmentation masks of much higher quality.

Weakly-Supervised VOS. Recent work on weakly-supervised VOS mainly focused on learning VOS only from unlabeled videos [20, 21, 23, 19, 34, 39, 23]. While being scientifically interesting, these approaches still leave a big gap to the results which can be achieved with full supervision. In our work, we make use of more supervision and get very close to the results of fully-supervised methods.

Reducing the Effort of Pixel-wise Annotations. There are many methods for semi-automatic image-level segmentation, which can be used to speed up the annotation process. For example, user input in the form of scribbles [29], bounding boxes [35], or clicks [36, 25, 16] can be used to interactively segment objects. Other examples are automatically predicted polygons [5, 1] or segments [4], which can be manipulated by the user. While speeding up the annotation for static images, most of these methods do not make use of existing bounding box level annotations and we found them often not to be robust enough to produce good results on out of domain data.

Recently, Voigtlaender *et al*. [32] introduced a semi-automatic annotation procedure to create dense segmentation labels for videos based on bounding box tracking annotations. Similar to our work, they use a convolutional network which converts bounding boxes into segmentation masks. However, they considered much smaller datasets and invested significant manual interaction in order to remove any errors in the annotations. Instead, we focus on a much larger dataset and allow some remaining errors in the annotations. This allows us to significantly reduce the amount of manual effort required, and we show that the remaining errors only have a minor impact on the final performance of models trained using these labels.

3. Reducing the Annotation Effort for VOS

Annotating dense segmentation labels for videos is extremely expensive. However, just annotating bounding boxes is much less effort and has already been done for more datasets like TAO [10]. In this paper, we evaluate which level of results can be achieved with a reduced set of annotations by automatically converting bounding boxes into segmentation masks. Furthermore, we investigate how the usefulness for training of automatically created pseudo-labels can be further improved.

3.1. Creating Pseudo-Labels from Boxes

When annotating segmentation labels for static images, there is much less redundancy, since there is a much larger variety between different images than between different frames of the same video, and segmentation labels for static image datasets like COCO [22] are readily available. We hypothesize that the information from already labeled static image datasets can be automatically transferred to videos for which only bounding boxes are labeled. In order to investigate how well this works, we adopt the powerful Box2Seg method for converting bounding boxes into segmentation masks [24].

Box2Seg. The Box2Seg network has also been used successfully in other works [24, 32, 33]. Box2Seg is a fully-convolutional network based on the DeepLabv3+ [7] segmentation architecture which uses an Xception-65 [8] backbone. The input to Box2Seg is an image together with a bounding box. The image is concatenated with a fourth input channel which encodes the bounding box as a mask. Afterwards, all four input channels are cropped to the box with an added margin, and then resized to a fixed size of 385×385 pixels. We train Box2Seg on COCO to predict the pixel-precise mask of an object given its bounding box.

Fine-Tuning. Just using a bounding box as input, Box2Seg already produces pseudo-labels of high quality as will be shown in Sec. 5. However, for some applications it might be necessary to get even better labels without resorting to extensive labeling of every frame. We can annotate a single mask (in a single frame) for each object and then use this mask to fine-tune Box2Seg on it for 300 steps before creating the masks for this object on the other frames. We hypothesize that this should achieve a good effort/quality trade-off. We evaluate this methodology on the YouTube-VOS 2018 [37] dataset. Here, on average 25 masks are annotated for each object. Hence, by annotating only a single mask, the effort is reduced 25-fold, and yet using the resulting labels to train, a result which is very close to using all annotations can be achieved (*c.f*. Sec. 5).

In the usual semi-supervised VOS task, always the first frame in which an object appears is given as annotated ground truth. However, using the first frame is a sub-

Figure 1. Qualitative example of automatically obtained pseudo labels on the YouTube-VOS training set. The first row shows the results produced by Box2Seg without using any manual mask annotations. The second row shows the result after fine-tuning on a single annotated frame per object. It can be seen that the segmentation quality improves significantly by only adding one mask per object manually. Note that this is a hard example and the mask quality looks better for most sequences.

optimal choice, and using instead the middle frame is a better heuristic [13]. Hence, for each object we take the middle frame from all frames in which this object appears. The improvement of the quality of pseudo-labels through fine-tuning is illustrated in Fig. 1.

3.2. Training using Pseudo-Labels

After obtaining the pseudo-labels from Box2Seg, we use them to train a VOS method on them. Here, we want to investigate what performance can be achieved by performing the training using the obtained pseudo-labels instead of the manually annotated ground truth.

As a case study, we use the extremely successful space-time memory network (STM-VOS) [27] architecture. STM-VOS uses two networks. The first network is a memory encoder, which takes an image together with a segmentation mask as input and produces key and value embeddings used as memory to segment further frames. This mask used by the memory encoder can be either given as first-frame ground truth, or it can be the prediction of STM-VOS for a previous frame. The second network consists of a query encoder, a space-time memory read block, and a decoder. The query encoder takes as input the current frame without a mask, and also produces key and value embeddings. The space-time memory read block then uses these embeddings for the current frame to produce feature maps by attending to the memory. The resulting feature maps are processed by the decoder to produce the segmentation masks for the current frame.

STM-VOS is first pre-trained on simulated data created from static image datasets, and afterwards the main training is done on a densely labeled video object segmentation dataset like YouTube-VOS. As baseline, we perform the main training of STM-VOS without any change in the training procedure, except for replacing the hand-labeled ground truth with the generated pseudo-labels. Although Box2Seg generally performs well, there will still be some errors in the pseudo-labels. For an effective training, it is important to limit the adverse effect of these errors. To this end, we evaluate two extensions for improving the results.

Filtering out Bad Masks. One possibility to deal with errors in the labels is to manually filter out bad masks. This does require manual effort, but is significantly cheaper than labeling all masks by hand. We evaluate how suitable this technique is by simulating this process on YouTube-VOS. To this end, we calculate the intersection-over-union (IoU) of the pseudo-label with the ground truth mask and mark its quality as bad if the IoU is below a certain threshold. In order to disregard the bad masks during training, in the pseudo-labels we remove this mask and instead fill the bounding box of the object with an ignore label, for which during training no loss is applied.

Robust Loss Function. An alternative to limit the influence of errors in the training annotations, is to use a more robust loss function. In the default setup, STM-VOS is trained by applying a pixel-wise cross entropy loss, *i.e.*

$$l_\theta(x, y) = -\log p_\theta(x, y), \tag{1}$$

where x is the input, y is the label for the considered pixel, and θ are the parameters to be optimized. For this loss function, the influence of a single pixel is unbounded and loss values can increase dramatically for wrongly-labeled data. In the context of learning classification models with label noise, Menon *et al*. [26] propose the partially Huberised cross entropy loss

$$\tilde{l}_\theta(x, y) = \begin{cases} -\tau \cdot p_\theta(x, y) + \log \tau + 1 & \text{if } p_\theta(x, y) \leq \frac{1}{\tau} \\ -\log p_\theta(x, y) & \text{else,} \end{cases} \tag{2}$$

which asymptotically saturates so that the influence of wrongly labeled data is bounded and the result for training

Dataset	Videos train	Videos val	Avg len.(s)	Objects / video	Ann. fps	Categories
DAVIS 2017	60	30	3.4	1.7	20	-
YouTube-VOS 2018	3,471	474	4.5	2.3	6	94
TAO (bound. boxes)	500	988	36.8	5.9	1	833
TAO-VOS	500	126	36.7	5.9	1	361

Table 1. Statistics of the most popular VOS datasets DAVIS and YouTube-VOS together with TAO (annotated only with bounding boxes) and the newly introduced TAO-VOS annotations. Sequences in TAO-VOS are significantly longer, cover more classes, and more objects per video than in previous VOS datasets. TAO-VOS has only 361 classes instead of 833, because we only annotated part of the validation set and we did not annotate the test set of TAO-VOS, and many classes only appear in the test set.

with noisy labels can be dramatically improved. The hyperparameter τ should be set according to the expected amount of noise in the labels. We adapt this loss to the task of VOS by applying it to every pixel individually with $\tau = 3$ and afterwards taking the average over spatial positions as usual.

4. The TAO-VOS Benchmark

Recently, the Tracking Any Object (TAO) [10] benchmark for multi-object tracking has been introduced. It consists of 500 videos for training, 988 videos for validation, and 1,419 videos for testing. TAO features 833 different classes, and contains on average 5.9 objects per video. To ensure a high diversity of the data, TAO is composed of videos obtained from seven different existing video datasets: Charades [30], ArgoVerse [6], HACS [42], AVA [14], YFCC100M [31], BDD [41], and LaSOT [11].

Tab. 1 shows statistics of the TAO dataset and compares it to the most important VOS datasets. It can be seen, that with 36.8 seconds, the average duration of the videos of TAO are much longer than in DAVIS 2017 [28] and YouTube-VOS [37]. Additionally, the average number of objects per video is much higher. DAVIS does not use predefined categories and the number of videos is very small. With 833 categories, TAO covers much more diversity than YouTube-VOS which features 94 classes. Moreover, for DAVIS and YouTube-VOS, objects almost always occupy a large fraction of the video, while for TAO new objects often appear and disappear.

Due to these unique challenges, TAO has a strong potential to serve as a new VOS benchmark. However, due to its large size, it was only annotated with bounding boxes. Furthermore, TAO is annotated at a lower frame-rate than DAVIS and YouTube-VOS, but this is not a problem since most modern VOS methods sample a few frames with a random time gap for each training step. Moreover, not all objects in all videos are annotated in TAO, however this is also not a problem for VOS, where the objects of interest are specified by their first-frame ground truth masks. In order to make this benchmark usable for VOS, we manually label the middle frame of each object of the training set with a mask, and then use Box2Seg with fine-tuning (*c.f.* Sec. 3.1) to generate masks for the whole training set of 500 sequences. Because TAO contains many more objects per sequence and is significantly more challenging than YouTube-VOS, the quality of the pseudo-labels is lower. Hence, in order to guarantee a high quality of the training set, we manually identified pseudo-labels with errors and added additional masks to fix the majority of these. First, for the middle frames, we manually created 5,666 masks. Afterwards, we manually added another 9,680 masks to guarantee a high quality of pseudo-labels. In total, 12,513 of the total 59,200 masks (21%) were annotated manually.

While the training set can be semi-automatically labeled, we argue that for evaluation, a definite ground truth without any mistakes is important. Hence, we select a representative subset (same amount from each sub-dataset) of 126 video sequences of the 988 validation videos and label each frame which has bounding box annotations by hand with masks to obtain the TAO-VOS validation set. In total, we annotated 14,987 masks for the validation set.

Tab. 1 shows statistics of the new TAO-VOS benchmark and Figures 2 and 3 show examples of the annotations on the training and the validation set, respectively.

5. Experiments

In the following, we perform experiments to validate the quality of the pseudo-labels, their effectiveness for training VOS models, and finally we present results of state-of-the-art VOS methods on the new TAO-VOS benchmark.

5.1. Quality of Pseudo-Labels

As our first experiment, we evaluate the quality of the pseudo-labels for the YouTube-VOS training set consisting of 3,471 sequences. Tab. 2 shows the mean IoU $\mathcal{J}$ together with the number of used manually annotated masks for different sets of labels. For a conventional VOS evaluation, also frames for which an object is not present are evaluated and the IoU is set to 1.0 if also the predicted mask is empty. Since we assume that the ground truth boxes are given, here we take the average only over frames in which an object is present, which leads to lower scores but gives a more intuitive understanding of the label quality.

Without annotating any masks, a mean IoU of 82.1 is reached. When fine-tuning using the middle frame, this result is improved by 5.5 percentage points at the cost of annotating only 6,459 masks. It can also be seen that using the middle frame for fine-tuning is slightly more effective than using the first frame. Fig. 1 shows a qualitative example of the label quality without and with fine-tuning using the middle frame.

Figure 2. Example annotations of the TAO-VOS training set. We show one frame of one example sequence for each sub-dataset.

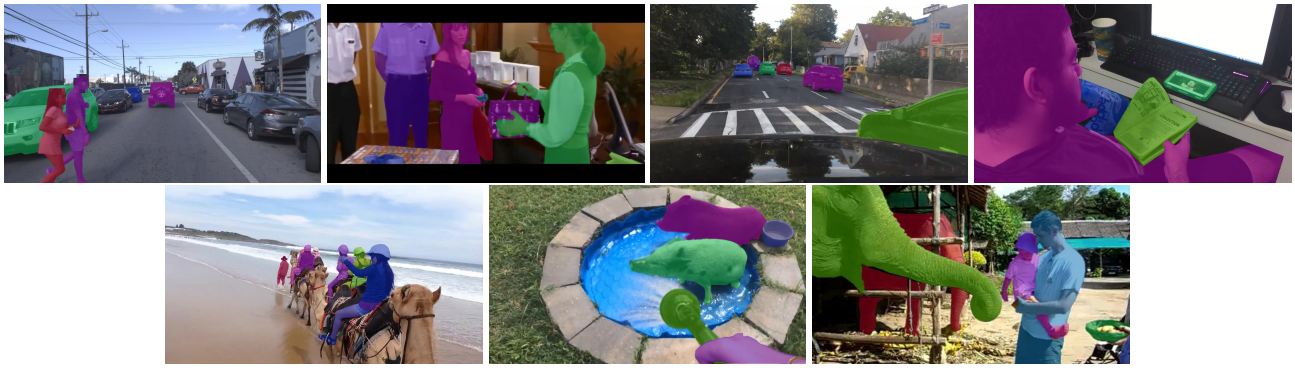

Figure 3. Example annotations of the TAO-VOS validation set. We show one frame of one example sequence for each sub-dataset.

Another way to evaluate the labels is to define a quality requirement for each mask, *e.g*. we require a mask to have at least 50% or 70% IoU with the ground truth mask and then we measure the fraction of automatically generated masks for which this quality is reached. The corresponding results are shown in Tab. 2 as recall@50% (recall at 50% IoU threshold) and as recall@70%. It can be seen that without using any manual labels, already 94.3% of the masks reach an IoU of at least 50% and 85.3% even reach an IoU of at least 70%. When fine-tuning using the middle frame, these values are largely improved to 97.9% for the 50% threshold and to 93.1% for the 70% threshold.

5.2. Training using Pseudo-Labels

Now we want to investigate how the label quality translates to the final VOS performance obtained when using these labels for training. Tab. 3 shows the results together with the number of used annotated masks on video datasets. For each configuration, first STM-VOS is pre-trained using static image datasets, and afterwards further trained on YouTube-VOS using the specified set of labels.

The results of the first three lines are directly copied from the STM-VOS paper [27], and indicate that training only using static images leads to poor results. Additionally, when adding the small but densely-annotated DAVIS 2017 training set, the results on the DAVIS 2017 validation set improve, but the results on the YouTube-VOS 2018 validation set degrade due to over-fitting.

For our own experiments, we use the code and the initialization checkpoint from static image pre-training provided by the authors. We then re-implemented the missing training code, and re-ran the training with full supervision and closely matched the results provided by the authors. We further show that the robust loss function does not lead to a significant change when applied with full supervision.

When training using the Box2Seg labels which require 0 annotated masks on video datasets, we achieve a $\mathcal{J}\&\mathcal{F}$ score of 74.9 on the DAVIS 2017 validation set and 74.3 on the YouTube-VOS 2018 validation set, respectively. This is only 3.9 percentage points less than our fully-supervised version on DAVIS and 5.0 percentage points less on YouTube-VOS, albeit the effort to annotate 159,976 segmentation masks was saved. Note that using the Box2Seg labels performs significantly better than both only doing static-image pre-training and to train using the fully-annotated DAVIS 2017 train set.

When manually filtering out bad masks (*c.f*. Sec. 3.2), this gap reduced further by 1.1 percentage points on DAVIS,

Labels	$\mathcal{J}$	recall@50%	recall@70%	Annotated masks (video)
Ground Truth	100.0	100.0	100.0	159,976
Box2Seg	82.1	94.3	85.3	0
Box2Seg + 1st frame FT.	86.7	97.4	91.8	6,459
Box2Seg + middle frame FT.	87.6	97.9	93.1	6,459

Table 2. Quality of pseudo-labels together with the number of used annotated ground truth masks on the YouTube-VOS training set. FT denotes fine-tuning, Annotated masks (video) refers to the number of used annotated masks on video datasets (excluding static image datasets), and recall@x% denotes recall (in percent) at an IoU threshold of x%.

Setup	Annotated Masks for Videos	DAVIS'17 val $\mathcal{J}\&\mathcal{F}$	YouTube-VOS'18 val $\mathcal{J}\&\mathcal{F}$
Fully-Supervised (YouTube-VOS) [27]	159,976	78.6	79.4
Fully-Supervised (DAVIS'17) [27]	9,627	71.6	56.3
Pre-training only [27]	0	60.0	69.1
Fully-Supervised (ours)	159,976	78.8	79.3
Fully-Supervised (ours) + Robust Loss	159,976	78.2 (-0.6)	79.8 (+0.5)
Box2Seg Labels	0	74.9 (-3.9)	74.3 (-5.0)
Box2Seg labels + Filtering	*	76.0 (-2.8)	74.6 (-4.7)
Box2Seg labels + Robust Loss	0	76.1 (-2.7)	76.2 (-3.1)
Box2Seg Fine-tuned labels	6,459	**77.7 (-1.1)**	77.3 (-2.0)
Box2Seg fine-tuned labels + Robust Loss	6,459	76.9 (-1.9)	**77.9 (-1.4)**

Table 3. Results of STM-VOS when training with different labels. Fully-Supervised (DAVIS'17): training on DAVIS 2017 instead of YouTube-VOS. Pre-training only: training only on static image datasets. Fully-Supervised (ours): our re-produced result using the code from STM-VOS [27]. Box2Seg labels + Filtering: using the pseudo-labels from Box2Seg and replacing low-quality masks with an ignore label (*c.f.* Sec. 3.2). *: indicates manual effort to assess the quality of the masks. Robust Loss: use of partially Huberised cross entropy loss (*c.f.* Sec. 3.2). Box2Seg fine-tuned labels: using the labels from Box2Seg which has been fine-tuned on the middle frame (*c.f.* Sec. 3.1).

and by 0.3 percentage points on YouTube-VOS. However, manual effort needs to be spent to assess the quality of the masks, which seems not worth it for the small improvement. We find that using the robust loss function (*c.f.* Sec. 3.2) instead is more effective (1.2 and 1.9 percentage points improvement for DAVIS and YouTube-VOS, respectively), and does not require any manual effort.

When annotating the 6,459 middle frame masks (roughly 4% of all masks) for fine-tuning Box2Seg, the gap to full supervision is greatly reduced to a small difference of 1.1 percentage points and 2.0 percentage points for DAVIS 2017 and YouTube-VOS, respectively. Using the labels obtained with fine-tuning, because of the higher label quality, the effect of the robust loss function is less significant. The robust loss slightly improves the results on YouTube-VOS, while slightly degrading the results on DAVIS.

In the supplemental material, we show that training using Box2Seg labels is also effective for the related task of video instance segmentation [38].

5.3. Comparison to Weakly-Supervised Methods

In addition to interpreting our approach as a method for reducing the annotation effort, the combination of pseudo-label generation by Box2Seg and then training STM-VOS on these labels can also be seen as a weakly-supervised method for VOS. In Tab. 4 we compare the results of this approach to other weakly-supervised VOS methods. We are only aware of two other methods, PReMVOS [24], and Siam R-CNN [33], which consider the training setup of only using bounding box annotations for videos, but making use of segmentation masks for static image datasets.

On DAVIS 2017, PReMVOS slightly outperforms our approach, but this comes at the cost of having to use four different neural networks and a run-time of more than 37 seconds per frame (compared to 0.32 seconds per frame for STM-VOS). For YouTube-VOS, PReMVOS has a significant gap of almost 8 percentage points to our approach. STM-VOS with Box2Seg labels (0 segmentation annotations for videos required), achieves a $\mathcal{J}\&\mathcal{F}$ score of 76.2, while the closest result is by Siam R-CNN [33] with 68.3.

We also compare to a number of approaches which use even weaker supervision: MuG [23] uses ImageNet classification labels and unlabeled videos, while MAST [19], UVC [21], and CorrFlow [20] are learned only from unlabeled videos. While these methods address an interesting research topic and achieve impressive results given that they do not need any annotations, the overall result quality is still far away from what can be achieved by including

Setup	Ann. Masks for Images	Ann. Boxes for Videos	Ann. Masks for Videos	DAVIS'17 val $\mathcal{J}\&\mathcal{F}$	YouTube-VOS'18 val $\mathcal{J}\&\mathcal{F}$
STM-VOS (Box2Seg Labels + Robust Loss)	✓	✓	0	76.1	76.2
STM-STM (Box2Seg Fine-tuned Labels)	✓	✓	6,459	77.7	77.3
PReMVOS [24]	✓	✓	0	77.8	66.9
Siam R-CNN [33]	✓	✓	0	70.6	68.3
MAST [19]	✗	✗	0	65.5	64.2
MuG [23]	✗	✗	0	56.1	-
UVC [21]	✗	✗	0	59.5	-
CorrFlow [20]	✗	✗	0	50.3	46.6

Table 4. Comparison of our results to other methods for weakly-supervised VOS. Here, Ann. Masks and Ann. Boxes denote the use of annotated masks and bounding boxes, respectively. Especially on YouTube-VOS, there is a big gap to the results of all previous methods.

Method	TAO-VOS $\mathcal{J}\&\mathcal{F}$	DAVIS'17 $\mathcal{J}\&\mathcal{F}$	YT-VOS'18 $\mathcal{J}\&\mathcal{F}$
STM [27]	63.8	81.8	79.4
CFBI [40]	66.3	81.9	81.4
LWL [12]	59.5	81.6	81.5

Table 5. Results on the validation sets of TAO-VOS, DAVIS 2017, and YouTube-VOS. The result for STM-VOS for DAVIS here includes fine-tuning on the DAVIS 2017 training set and all results on TAO-VOS include fine-tuning on the TAO-VOS training set.

Method	TAO-VOS val $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
STM [27] (YouTube-VOS)	62.3	59.9	64.7
+ TAO-VOS fine-tuning	63.8 (+1.5)	61.5	66.1
CFBI [40] (YouTube-VOS)	63.9	61.9	66.0
+ TAO-VOS fine-tuning	66.3 (+2.4)	63.8	68.8
LWL [12] (YouTube-VOS)	55.9	53.0	58.7
+ TAO-VOS fine-tuning	59.5 (+3.6)	56.7	62.4

Table 6. Effect of fine-tuning on the TAO-VOS training set.

static-image training data with segmentation labels.

5.4. TAO-VOS

Next we evaluate recent VOS methods on the new TAO-VOS benchmark. We consider three methods that together cover the current state-of-the-art: STM-VOS [27], CFBI [40], and LWL [12]. CFBI employs a global and a local matching strategy to match the pixels of the current frame to pixels of other frames in order to transfer information which is used as internal guidance of the network to produce masks. Additionally, a special handling of the background and an instance-level attention mechanism are used. LWL [40] uses meta-learning to learn a target model of the object of interest. The target model predicts an encoding of the target mask, which is used by a decoder to produce segmentation masks. All three approaches are state-of-the-art VOS methods and achieve excellent and very similar performance numbers on DAVIS and YouTube-VOS.

The TAO-VOS training set only consists of 500 sequences which is still much smaller than YouTube-VOS with 3,471 sequences. Hence, we initialize from YouTube-VOS pre-training and use TAO-VOS for fine-tuning (see supplemental material for implementation details).

The TAO-VOS validation set is only annotated at 1 frame per second (FPS), while the video data is available at around 30 FPS. Similar to the evaluation protocol for YouTube-VOS, an algorithm can either run only on the annotated frames, or it can use all frames. Intuitively, tracking objects at a higher frame-rate should be easier. For each method we tried both variants and report the best of the two results. Interestingly, we found that only LWL can benefit from the higher frame rate, while STM and CFBI produces better results when run only on the annotated frames (see supplemental material for details).

Tab. 5 shows the results on TAO-VOS after fine-tuning, and we also show the results of the original methods on the validation sets of DAVIS 2017 and YouTube-VOS 2018 as a reference. It can be noted that the absolute performance on TAO-VOS is much lower than on DAVIS and YouTube-VOS. Furthermore, all three methods perform very similar on DAVIS and YouTube-VOS, while on TAO-VOS the result for LWL is significantly worse than the results of the two other methods. These results show that TAO-VOS is much more challenging than DAVIS and YouTube-VOS, and that it can further reveal differences between methods which were not visible on existing datasets. The best result of 66.3 $\mathcal{J}\&\mathcal{F}$ is relatively low and leaves much room for improvement by further methods, while the performance on DAVIS and YouTube-VOS shows strong signs of saturating.

In Tab. 6, we analyze the effect of fine-tuning on the TAO-VOS training set. It can be seen that all three methods benefit from fine-tuning with the largest improvement of 3.6 percentage points achieved by LWL.

Fig. 4 shows qualitative results on two example sequences. It can be seen that both considered sequences are

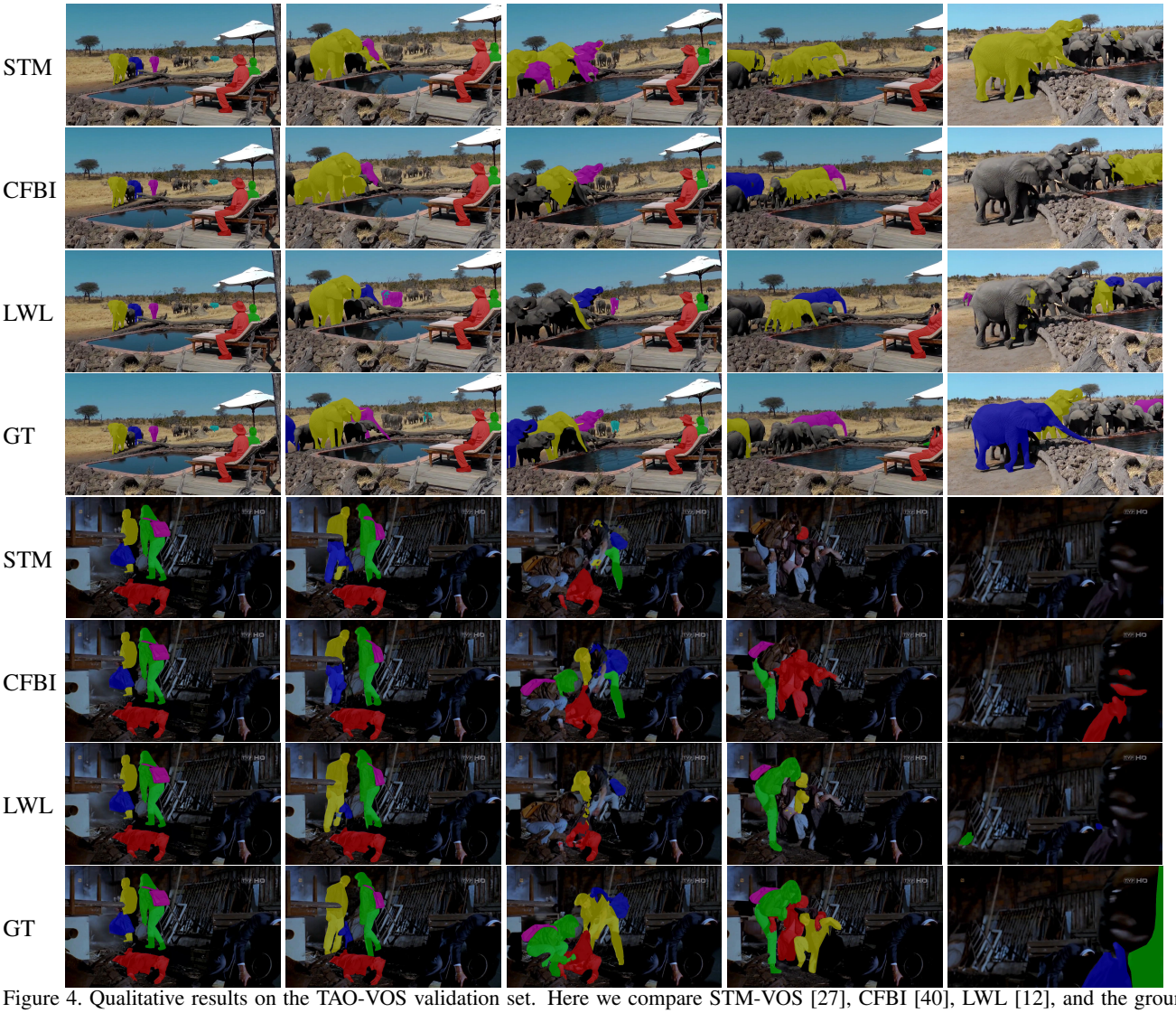

Figure 4. Qualitative results on the TAO-VOS validation set. Here we compare STM-VOS [27], CFBI [40], LWL [12], and the ground truth (GT). For the two considered hard sequences from TAO-VOS, all considered VOS methods fail to segment the objects accurately after some time.

very hard with many objects and that all three considered methods lose track of all objects after some time. Again, this highlights that TAO-VOS as a benchmark has strong potential to enable further progress in VOS.

6. Conclusion

We started with the observation that the progress of VOS is strongly driven by the availability of VOS datasets, and that current VOS datasets are still relatively small and do not cover diverse, long, and challenging enough videos. Subsequently, we identified the high manual effort of dense pixel-wise annotations as a major bottleneck for further progress in VOS. Towards a solution, we show that dense labels might not be necessary, and only bounding boxes plus few or no segmentation annotations are enough to create pseudo-labels from which VOS methods can learn almost as much as from costly dense annotations. As a proof of concept, we labeled the TAO-VOS training set semi-automatically with high quality. In contrast to VOS performance saturating on current VOS benchmarks, the challenging TAO-VOS validation set reveals differences between and shortcomings of state-of-the-art VOS methods. We hope that our work inspires the creation of even larger VOS datasets which can be labeled with less effort.

Acknowledgements: This project has been funded, in parts, by ERC Consolidator Grant DeeViSe (ERC-2017-COG-773161), a Google Faculty Research Award, NSFC project Grant No. U1833101, Shenzhen Science and Technologies project under Grant No. JCYJ20190809172201639 and the Joint Research Center of Tencent and Tsinghua. The authors would like to thank Berin Gnana and Sabarinath Mahadevan for helpful discussions.

References

[1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *CVPR*, 2018.

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019.

[3] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018.

[4] Misha Andriluka, Jasper Uijlings, and Vittorio Ferrari. Fluid annotation: a human-machine collaboration interface for full image annotation. *arXiv preprint arXiv:1806.07527*, 2018.

[5] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *CVPR*, 2017.

[6] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

[9] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

[10] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. *arXiv preprint arXiv:2005.10356*, 2020.

[11] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling. LaSOT: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019.

[12] Bhat Goutam, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *ECCV*, 2020.

[13] Brent A. Griffin and Jason J. Corso. Bubblenets: Learning to select the guidance frame in video object segmentation by deep sorting frames. In *CVPR*, 2019.

[14] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018.

[15] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In *NeurIPS*, 2019.

[16] Suyog Dutt Jain and Kristen Grauman. Click carving: Interactive object segmentation in images and videos with point clicks. *IJCV*, 2019.

[17] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017.

[18] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.

[19] Zihang Lai, Erika Lu, and Weidi Xie. MAST: A memory-augmented self-supervised tracker. In *CVPR*, 2020.

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

[21] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *NeurIPS*, 2019.

[22] T-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[23] Xiankai Lu, Wenguan Wang, Jianbing Shen, Yu-Wing Tai, David J. Crandall, and Steven C. H. Hoi. Learning video object segmentation from unlabeled videos. In *CVPR*, 2020.

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

[25] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively trained interactive segmentation. In *BMVC*, 2018.

[26] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2020.

[27] S. Wug Oh, J.-Y. Lee, N. Xu, and S. Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019.

[28] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[29] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. GrabCut: Interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.

[30] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016.

[31] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016.

[32] Paul Voigtlaender, Michael Krause, Aljoša Ošep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *CVPR*, 2019.

[33] Paul Voigtlaender, Jonathon Luiten, Philip H. S. Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *CVPR*, 2020.

[34] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J. Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, 2019.

[35] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep grabcut for object selection. In *BMVC*, 2017.

[36] Ning Xu, Brian L. Price, Scott Cohen, Jimei Yang, and Thomas S. Huang. Deep interactive object selection. In *CVPR*, 2016.

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

[38] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019.

[39] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip H. S. Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, 2019.

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

[41] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020.

[42] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *ICCV*, 2019.