ELEKTROTEHNIŠKI VESTNIK 90(4): 147–158, 2023
ORIGINAL SCIENTIFIC PAPER

Guided Video Object Segmentation by Tracking

Jer Pelhan[1,†], **Matej Kristan**[1], **Alan Lukežič**[1], **Jiri Matas**[2], **Luka Čehovin Zajc**[1],

[1] *Faculty of Computer and Information Science, University of Ljubljana,
Večna pot 113, 1000 Ljubljana, Slovenia*
[2] *The Center for Machine Perception, Czech Technical University,
166 36 Prague 6, Prague, Czech Republic*

[†] *E-mail: jer.pelhan@fri.uni-lj.si*

Abstract. The paper presents *Guided video object segmentation by tracking (gVOST)* method for a human-in-the-loop video object segmentation which significantly reduces the manual annotation effort. The method is designed for an interactive object segmentation in a wide range of videos with a minimal user input. User to iteratively selects and annotates a small set of anchor frames by just a few clicks on the object border. The segmentation then is propagated to intermediate frames. Experiments show that gVOST performs well on diverse and challenging videos used in visual object tracking (VOT2020 dataset) where it achieves an IoU of 73% at only 5% of the user annotated frames. This shortens the annotation time by 98% compared to the brute force approach. gVOST outperforms the state-of-the-art interactive video object segmentation methods on the VOT2020 dataset and performs comparably on a less diverse DAVIS video object segmentation dataset.

Keywords: convolutional neural network, video object segmentation, video object tracking, user interaction, data annotation

Vodenje segmentacije objektov v videoposnetku s sledenjem

V članku predlagamo *s sledenjem vodeno segmentacijo objekta v video posnetku*, metodo za interaktivno segmentacijo objekta v video posnetku. Metoda bistveno zmanjvsa delo pri procesu dodajanja natančnih segmentacij objektov. Zasnovana je za interaktivno segmentacijo objektov v vstevilnih videoposnetkih z minimalnim vnosom uporabnikov. Od uporabnika zahteva iterativno izbiro in označevanje majhnega nabora sidrnih okvirjev s samo nekaj kliki na meji objekta. Nato se segmentacija razvsiri na vmesne okvirje. Poskusi kažejo, da metoda dosega vrhunsko zmogljivost pri raznolikih in zahtevnih videoposnetkih, ki se uporabljajo pri vizualnem sledenju predmetom (nabor podatkov VOT2020), kjer doseže IoU 73 % pri samo 5 % uporabnivsko označenih okvirjev. To dejansko skrajša čas označevanja za 98 % v primerjavi z naivnim pristopom. Predlagana metoda prekavsa najsodobnejvse interaktivne metode segmentacije video objektov na prej omenjenem naboru podatkov VOT2020 in primerljivo deluje na manj raznoliki zbirki podatkov za segmentacijo objektov v videoposnetkih DAVIS.

1 INTRODUCTION

High-quality annotations are crucial in the development of modern computer vision methods. While they are traditionally important for objective evaluation, they have recently become the driving force for training ever-advancing deep learning models. Per-pixel segmentation is fundamental for various computer vision tasks, like object detection [8], video editing [24], surveillance [6] and autonomous driving [16]. But obtaining large manually annotated training and testing datasets, especially per-frame segmented videos, is time-consuming, error-prone and costly.

A number of semi-automatic user-guided segmentation methods have been proposed to address the need for accurately segmented objects in videos [10, 18, 3]. These methods, commonly referred to as video object segmentation (VOS) methods, focus on accurate segmentation in high-resolution videos. On the other hand, even the state-of-the-art video object segmentation methods are designed for segmentation over relatively short video sequences and lack robustness for small fast-moving objects that substantially change the appearance and low-quality videos with varying lighting conditions. This reduces their effectiveness and generality. Object generality is particularly emphasized in the field of *general visual object tracking*. An object tracking algorithm has to predict the position of the target in a sequence given only its location in the first frame. The tracker has to be robust over a range of objects and work in a variety of scenes. While general object tracking traditionally considers target locations as bounding boxes, there has been a recent shift towards complete target segmentation. For example, the major tracking challenge, VOT [11] has abandoned bounding boxes in short-term tracking category and now requires trackers

Received 29 June 2023
Accepted 19 August 2023

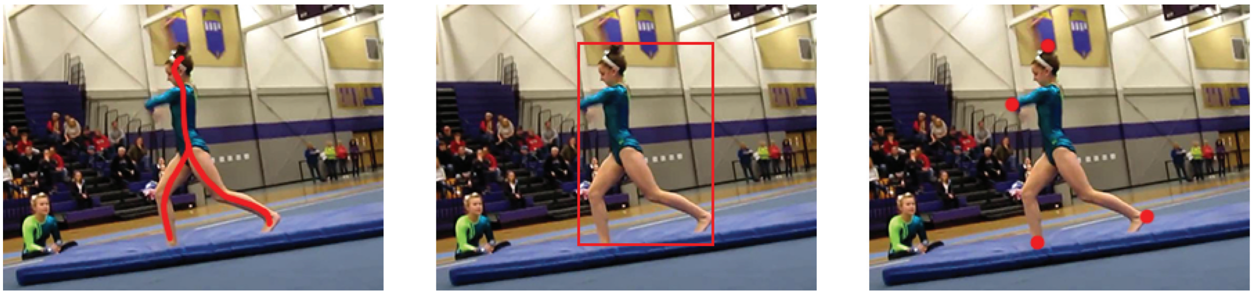


Figure 1. Overview of the inputs for different interactive image segmentation methods. From left to right: scribbles for [3, 12, 10, 18], bounding box for [27] and extreme points for [15].

to output segmentation masks. The most recent VOT challenge also demonstrates that general object trackers cope much better with challenging targets in long videos than video object segmentation algorithms.

In the paper, we address the issue of the video object dataset annotation. We draw our inspiration from the development in visual object tracking and propose an interactive method for a fast and accurate video object segmentation – gVOST (guided video object segmentation by tracking). gVOST uses the state-of-the-art single frame segmentation that allows user to accurately segment an object with a few clicks. This is combined with recent advances in VOT that enable the propagation of the mask from individual keyframes to the remaining frames. We also describe a method that selects the best mask proposal when there are several masks avaliable. Our experimental analysis shows that gVOST performs well over a range of videos and outperforms recent state-of-the-art video object segmentation methods on the most challenging ones. Furthermore, gVOST is integrated in an application for video annotation. The application is positively tested by independent researchers for annotation of tracking videos.

2 Related work

The user-guided video object segmentation methods can be roughly divided according to the type of the user-involvement into (a) semi-supervised and (b) interactive video object segmentation. Semi-supervised video object segmentation methods predict segmentation masks for the entire video based on the segmentation mask of the first frame. Many methods [26, 1] segment the target based on the features extracted in the first frame. Tomakov *et al.* [25] combine the motion data and an appearance embedding to maintain the state of object from frame to frame. [22] proposes a three-step approach, where multiple sequence-specific region proposals are generated for each video frame based on the mask in the first frame. A segment proposal tracking algorithm is used to label regions of query objects, which is followed by spatial refinement. These methods do not support any feedback refinement loop which makes them less suitable for an interactive video segmentation.

An interactive video annotation [2] extends the user control over segmentation as a feedback loop. Most of the state-of-the-art interactive VOS methods [18, 10, 21] are based on two steps: (i) interaction or annotation and (ii) propagation or transfer. In the interaction step, the user approximated segmentation masks are used from multiple frames and this information is aggregated into an appearance model. In [18, 10] the propagation algorithm predicts masks over empty frames using an appearance model. In [3], visual tracking is used to constrain segmentation to tracker-predicted bounding boxes whereas, our method fully integrates the segmentation tracker into the segmentation pipeline.

The user interaction with the segmentation method must be carefully designed. Since a manual mask annotation takes approximately 79 seconds per image [15], image segmentation techniques are used in the interactive VOT methods (Fig. 1). Most methods [10, 18, 3] use scribbles, i.e., a set of curves roughly covering the object area. A faster, but less accurate alternative, is initialization using bounding boxes [29, 27]. Unlike our method which requires the user to click a set of object boundary points, akin to [15, 13]. This kind of input requires less user effort as scribbles and avoids the cumbersome process of drawing accurate bounding-boxes.

3 Guided Video Object Segmentation by Tracking

Video object segmentation can be formally defined as a process that transforms a sequence of images $\{\mathbf{I}_0, ...\mathbf{I}_N\}$ into an equally long sequence of binary masks $\{\mathbf{M}_0, ...\mathbf{M}_N\}$, each corresponding to one input image. An object of interest is defined externally, either as a complete segmentation in a single frame or in an interaction with the user. A richer interaction leads to overall better results, but at the cost of the user engagement. It is therefore crucial to make a compromise between the user engagement and the segmentation quality.

The balance is ensured by minimizing the user work in two ways: (1) structural properties of natural images are used to efficiently segment individual images with a limited amount of the user input, and (2) a temporal consistency in videos to efficiently propagate masks between provided key frames, thus further reducing interaction where possible. Our method works in two steps repeated until the segmentation masks are sufficiently accurate.

The first step is the *anchor selection and segmentation step* in which the user selects frames that act as a guidance and annotates the masks for key frames in an efficient manner. The set of anchor indices is denoted as follows:

$$A = \{a_0, ...a_K \mid 0 \leq a_0 < a_1 < ... < a_{K-1} \leq N\}, \tag{1}$$

where K is the number of the anchor frames (minimally one is required). In the second step, masks on non-annotated frames are then predicted. They are predicted with *segmentation propagation by tracking*. The segmentation masks are propagated from anchors to the remaining frames using a visual object tracker capable of producing a segmentation output. The idea is illustrated in Figure 2. The tracker is run from one anchor frame to the other (or to the beginning or end of the sequence). When predicting the mask on a frame between two anchor frames, the tracker is initialized with the data from both anchors.

Each frame between the two anchors thus receives two mask predictions denoted as $\mathbf{P}_i^F$ (forward tracking prediction) and $\mathbf{P}_i^B$ (backward tracking prediction). A naive selection choosing the segmentation that is propagated from the closest anchor does not perform well if the object gets occluded or some other phenomenon causes a tracker failure. Therefore, a mask selection strategy based on an *external* and *internal* supervision is proposed.

The external supervision is based on the observation that the quality of the predicted mask at the end of the interval, i.e., the next anchor frame, reflects the accumulated error and thus the quality of the predictions throughout the interval. The quality is formalized as a J&F mask overlap score [20] between the predicted segmentation and the user provided segmentation. Since the score is calculated at the end of the interval, it is constant for all the frames within the interval. To determine the per-frame score a tracker-specific internal supervision is made based on the internal localization confidence for a given frame (the specific implementation is described in Section 3.1). The final score for the forward prediction, ($\mathbf{P}_i^F$), with an internal tracker confidence (C_i^F) at frame $i \in (a_s, a_e); e = s + 1$ is defined as

$$S_i^F = C_i^F \; \frac{\exp[\Lambda(\mathbf{P}_{a_e}^F, \mathbf{M}_{a_e})]}{\exp[\Lambda(\mathbf{P}_{a_e}^F, \mathbf{M}_{a_e})] + \exp[\Lambda(\mathbf{P}_{a_s}^B, \mathbf{M}_{a_s})]}, \tag{2}$$

where $\Lambda(\cdot, \cdot)$ is the segmentation similarity function, $P_{a_e}^F$ is the forward prediction (mask) in the end frame for the interval and $P_{a_s}^B$ is the backward prediction for the start frame of the interval. As the backward prediction score (S_i^B) is analogous, simply switching forward and backward elements of the equation is required. The final segmentation mask for a frame is selected as the prediction with the larger score, i.e.,

$$\mathbf{M}_i = \begin{cases} \mathbf{P}_i^F & \text{if } S_i^F > S_i^B, \\ \mathbf{P}_i^B & \text{otherwise.} \end{cases} \tag{3}$$

Our video object segmentation framework leverages two state-of-the-art methods, i.e. the interactive object image segmentation method for fast anchor frame segmentation and the visual object tracking method for segmentation propagation.

3.1 Anchor segmentation

DEXTR [15] is utilized for the interactive object image segmentation. It accepts at least four extreme points (right-most, left-most, top and bottom) of the object of interest as the input. To get a more accurate mask estimation more points can be added at the edge of the object (see Fig. 3). The segmentation mask prediction is obtained by encoding border points to a distance heatmap, concatenating the heatmap with a color information into four-channel image and passing it through a pre-trained fully convolutional auto-encoder network based on a Deeplab-v2 [23] architecture.

3.2 Segmentation propagation by tracking

Our segmentation propagation algorithm is based on the recent visual object tracker D3S [14]. It is one of the first successful deep learning segmentation-based visual object trackers that surpasses its traditional bounding box counterparts. D3S outputs the segmentation based on the information from two different target visual models – GIM and GEM. The *geometrically invariant model* (GIM) is based on a deep feature comparison* from a query frame to the target and background set of features obtained from an initialization frame. The background and foreground similarities of each feature are obtained as average of top K similarities to features extracted in the initialization phase allowing to build a segmentation probability map. The second model, *geometrically constrained model* (GEM), encodes target location provided by an adaptive deep discriminative correlation filter [4] as unimodal probability map. The outputs of both models are concatenated and upscaled

*D3S uses features extracted with a ResNet50 backbone [7].

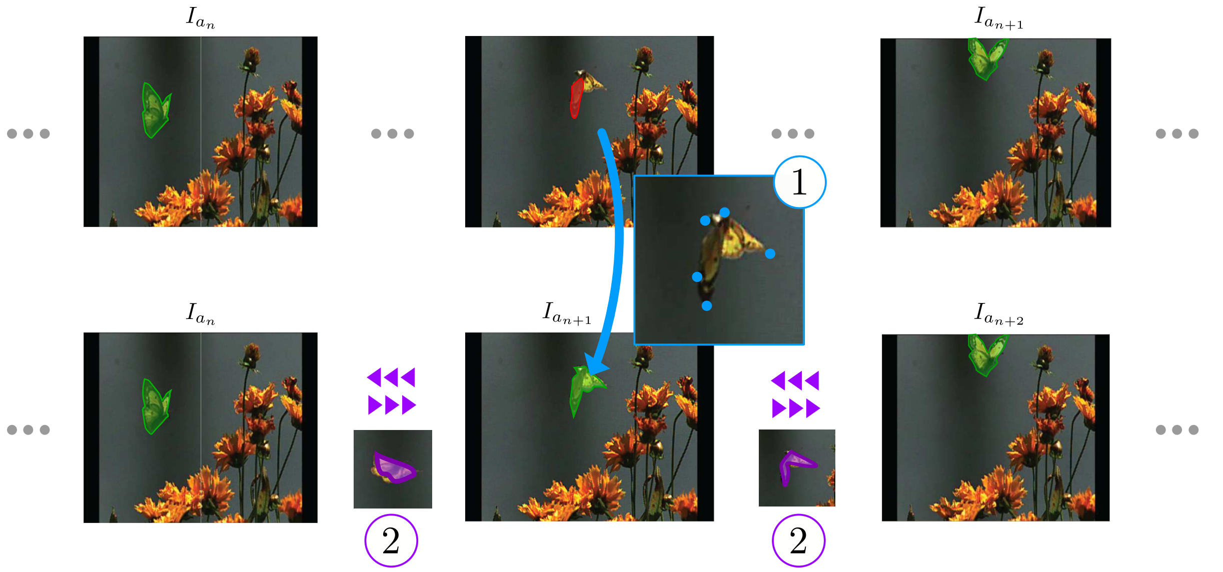


Figure 2. Single iteration of the interaction loop. (1) - user interaction, a frame between two anchors with a bad automatic segmentation is selected, a new mask is inserted by providing border points. (2) - automatic propagation, a new anchor is inserted and the masks for the frames in intervals connected to the new anchor are updated using an automatic propagation. Both steps are repeated until the resulting segmentation is satisfactory.

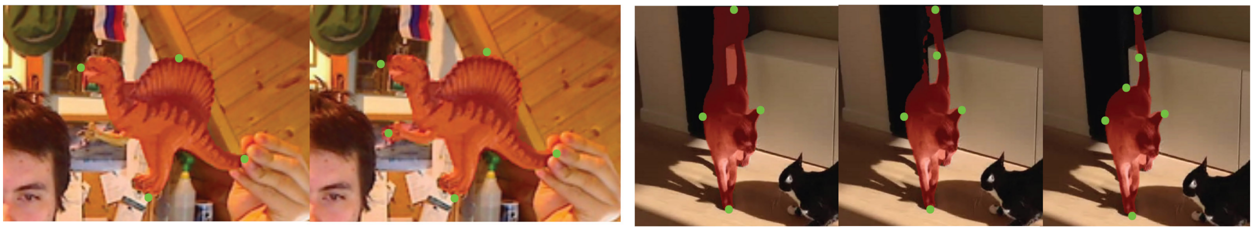


Figure 3. When the desired image segmentation accuracy is not achieved, additional points at the border of the object are added as an input to improve the segmentation.

with a refinement decoder network to a segmentation mask of the input size. D3S is originally initialized on a single (starting) frame. It is modified to use initializations from several frames, i.e., from the interval start and end anchor frame. The maximum correlation response value from GEM is used as the internal confidence score $C_i^{\cdot}$ required by the mask selection score (Equation 2).

4 Experiments

Several proposed video object segmentation datasets [2, 28] focus on accurate segmentation in short sequences with large objects, without the evaluation of their robustness in more difficult setups. To account for this, we consider a recent visual tracking performance evaluation dataset, the *Visual Object Tracking* (VOT2020) dataset [11], as our main evaluation dataset. To test the generalization capabilities, gVOST is on a well known visual object segmentation dataset – the *Densely Annotated Video Segmentation* (DAVIS2017) dataset [20]. Which holds a interactive video segmentation challenge allowing comparison with the most state-of-the-art interactive video object segmentation methods. Our method is compared with two state-of-the-art interactive video object segmentation methods, the winner of the DAVIS2020 interactive challenge IVOS [10] and the winner of the DAVIS2018 interactive challenge IVS [18].

4.1 Implementation Details

Our guided video object segmentation by tracking method (gVOST) is implemented in Python using PyTorch library*. A pre-trained *DEXTR* model [15], trained on PASCAL2012 [5] segmentation dataset is used. In the DAVIS2017 experiments, presented in Section 4.3, we additionally fine-tune the model on the training part of the dataset [20] for 50 epochs with learning rate of $1e{-}8$, momentum of 0.9 and weight decay of $5e{-}4$ with the batch size of five samples.

*https://pytorch.org/

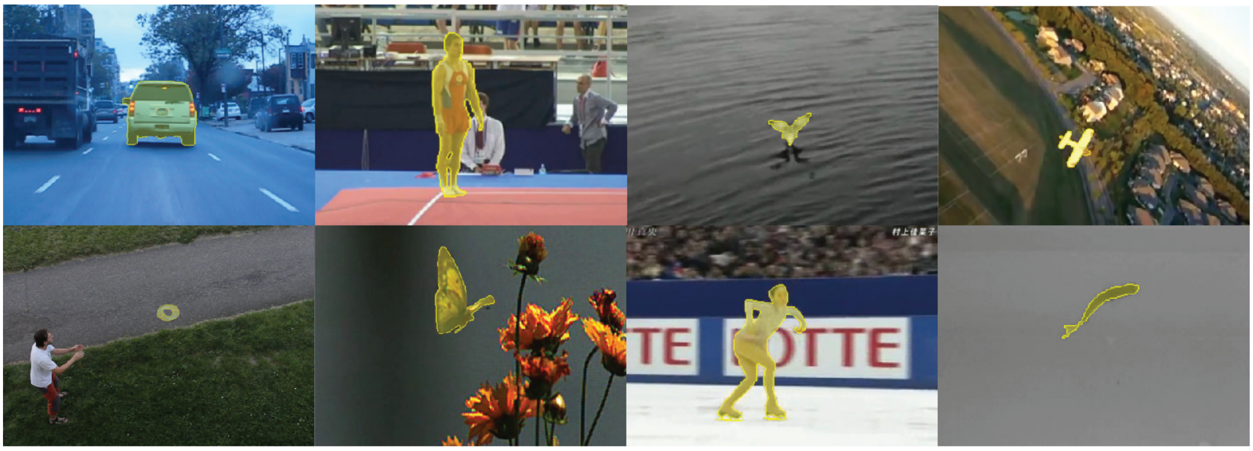


Figure 4. Examples of object segmentation masks carefully annotated by annotators in the user study.

To fully utilize the user annotations, the mask propagation module, D3S, is fine-tuned to the user-annotated anchors during an interactive segmentation for 3 epochs with 50 iterations and $1e{-}4$ initial learning rate with 0.2 per-epoch reduction.

4.2 Evaluation on the VOT2020 dataset

The VOT2020 dataset is used for our primary evaluation. This is the first visual object tracking dataset that contains the ground truth in a form of binary segmentation masks generated by experienced human annotators. The dataset contains a diverse set of 60 video sequences of average length 300 frames, which is several times longer than a typical video object segmentation dataset sequence (approximately 70 frames).

4.2.1 The annotation accuracy: General video sequences included in tracking datasets are quite challenging even for human annotators as the resolution tends to be modest and the target object may be blurred due to the fast motion, etc. This leads to a certain degree of annotation ambiguity. To estimate the achievable segmentation consensus, six volunteers were asked to carefully annotate objects on selected frames twice in a row. We have selected a set of 8 diverse frames from different sequences (see Figure 4) to capture the variability of objects, while keeping the cumulative effort of volunteers reasonably low. In this way we obtained 12 segmentation masks for each frame and 96 segmentation masks for the entire set. A visual evaluation shows that all segmentation masks were of the highest quality. Despite some minor differences between the masks for the same object, all the masks should be considered as the ground truth and their variation specifies the level under which the differences should be considered negligible for practical evaluation.

The VOT2020 ground truth is compared with the 96 masks obtained in the survey to estimate the level of segmentation ambiguity, i.e., the performance measure bound beyond which all alternative segmentation masks should be considered as equivalent. The average overlap for IoU and J&F is 0.84 and 0.89, respectively. Any overlap exceeding $\rho_{\mathrm{VOT}} = \mu_{\mathrm{VOT}} - \sigma_{\mathrm{VOT}}$ (σ_{VOT} being the standard deviation over the distribution of differences) is considered to be accurate beyond the annotation noise. Thus $\rho_{\mathrm{VOT}}^{\mathrm{IoU}} = 0.76$ and $\rho_{\mathrm{VOT}}^{\mathrm{J\&F}} = 0.82$ are referred to as the *VOT annotation accuracy bound*.

4.2.2 The evaluation protocol: An automatic evaluation protocol is required for a systematic, repeatable and reproducible analysis [20]. The DAVIS protocol, described in Section 4.3.1, iteratively prompts the user to annotate a single frame. Since the visual object tracking sequences are longer and more challenging than the video object segmentation sequences, we propose a modified protocol that more faithfully reflects practical interaction scenario. Anchors are selected at every Δth frame and segmented using a simulated user input derived from the ground truth. The algorithm then propagates the masks to other frames. A new anchor is added at the frame with the worst segmentation and the process is repeated. The performance is measured using the intersection over union (IoU) as well as the J&F measure that explicitly emphasises the accuracy of the segmentation mask at the border of the object [19].

The user interaction in the DEXTR control points input is simulated by inferring points from the ground truth for the corresponding frame. Four extreme points are first calculated from the ground truth segmentation mask. If needed, an additional point is added along the contour where the estimated mask deviates most from the ground truth. This process is limited to eight points, afterwards the mask with the largest J&F score is chosen as the final user-selected anchor mask. The reference methods [10, 18] used in our evaluation require scribbles for initialization. The scribbles generated for these methods use the standard method from the DAVIS interactive challenge toolbox [2].

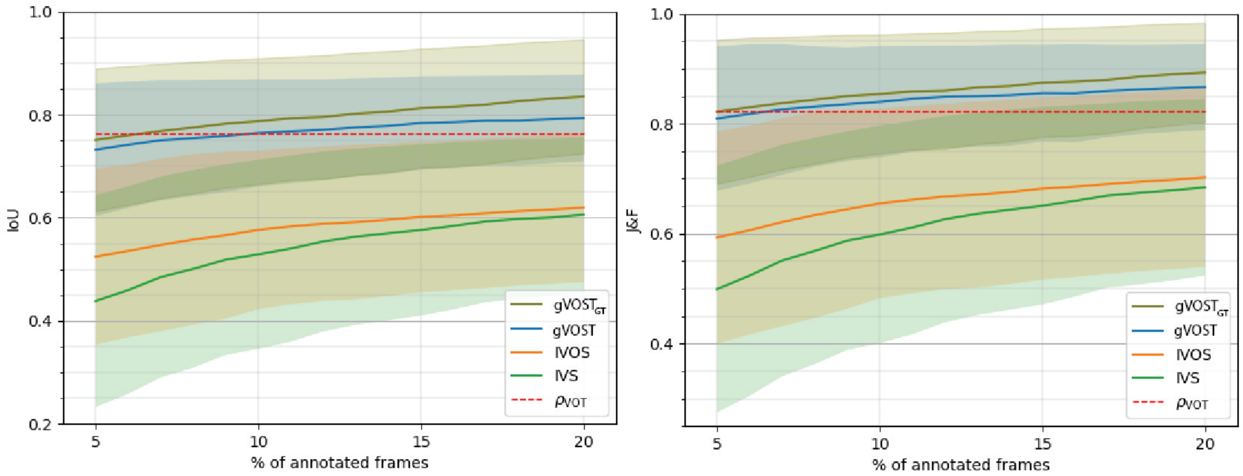


Figure 5. Segmentation accuracy on VOT2020 with respect to the percentage of annotated frames. gVOST surpasses the VOT2020 accuracy bound ρ_{VOT} at 10% of all frames annotated, significantly outperforming the competing methods.

4.2.3 Quantitative analysis: The performance on VOT2020 with respect to the percentage of annotated frames in a sequence is shown in Figure 5 and Table 1. On average, gVOST requires seven clicks per an anchor mask. At already 5% of the manually annotated frames, gVOST achieves on average 0.732 IoU, thus outperforming IVOS [10] and IVS [18] by 40% and 67%, respectively. gVOST achieves the VOT IoU annotation accuracy $\rho_{\mathrm{VOT}}^{\mathrm{IoU}}$ bound at annotating only 9% of all frames (0.76 IoU) and reaches the J&F VOT annotation accuracy boundary $\rho_{\mathrm{VOT}}^{\mathrm{J\&F}}$ at 7% annotated frames (0.82 J&F). Neither IVS nor IVOS comes close to the annotation accuracy bound within 20% of the annotated frames (every fifth frame).

Ω	Method	IoU	J&F
5%	IVS [18]	0.438	0.500
	IVOS [10]	0.524	0.592
	gVOST	**0.732**	**0.810**
10%	IVS [18]	0.529	0.598
	IVOS [10]	0.576	0.655
	gVOST	**0.764**	**0.842**
20%	IVS [18]	0.606	0.684
	IVOS [10]	0.620	0.702
	gVOST	**0.793**	**0.869**

Table 1. Segmentation accuracy on VOT2020 with respect to the percentage of the annotated frames – Ω. Best results are boldfaced.

According to the estimates [15], a manual segmentation of all frames in a typical tracking sequence (300 frames) takes 6 hours of constant work. Our analysis shows that gVOST requires user input on approximately every 17th frame to segment all the frames in the video with the masks that exceed the VOT annotation accuracy ρ_{VOT}. This means that gVOST reduces the annotation time by 94%, i.e. to merely 25 minutes. When using the DEXTR-based initialization with seven clicks, only every 10th frame needs to be annotated to segment a video with an equal quality. It takes ten seconds to click seven points on an object, thus the required video annotation time is reduced to mere five minutes. The conclusion is that gVOST decreases the user work for visual object tracking domain videos by 98%.

4.2.4 Ablation analysis: The contribution of different parts of gVOST is analyzed in an ablation study. The study is conducted on a subset of eight sequences, selected from the VOT2020 dataset[†]. We have compared the influence of fine-tuning the D3S tracker during the annotation process (gVOST$_{\mathrm{NFT}}$ - no fine-tuning) and the influence of initializing the tracker with the data from the anchors at the both sides of the interval (gVOST$_{\mathrm{N2}}$ - no second anchor) with the main method (gVOST). We have also included a version of the method that uses ground truth segmentation directly in anchors to study the influence of the DEXTR-based anchor initialization compared to inputting the segmentation mask manually (gVOST$_{\mathrm{GT}}$).

The results, given in Table 2, show that additional information about the object indeed improves the performance. The fine-tuning of the mask decoder adapts it to the current sequence, thus increasing the accuracy, resulting in 5% improvement. The influence of using two frames in the tracker initialization is less noticeable,

[†] The selected sequences are the same as were used as a source of the frames for the segmentation accuracy bound, shown in Figure 4.

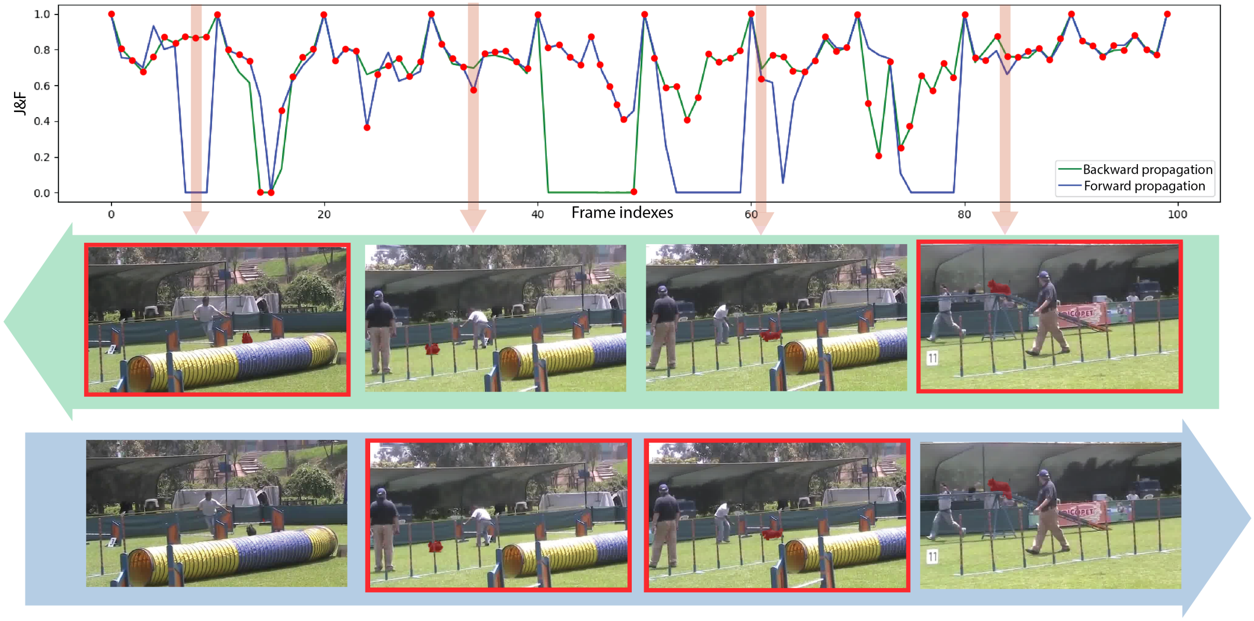


Figure 6. J&F overlap of the backward and forward-propagated masks from anchors at each frame. The forward and backward propagated masks for selected frames are shown below the plot. Red dots on the plot and red squares around the frames indicate the chosen mask from two possibilities.

	5%	10%	20%
gVOST$_{\text{NFT}}$	73.6/83.0	75.9/85.2	77.4/86.5
gVOST$_{\text{N2}}$	77.9/86.7	78.2/87.4	79.7/88.5
gVOST	78.0/86.8	79.5/88.3	80.5/89.1
gVOST$_{\text{GT}}$	80.4/88.0	82.7/89.9	83.6/92.3

Table 2. Ablation analysis results for eight sequences selected from the VOT2020 dataset with respect to the percentage of the annotated frames in the sequence. The performance is reported as an mIoU - J&F pair.

adding only 1%, but a detailed inspection showed that it contributes to the tracker robustness in the more challenging intervals. Using the ground truth mask to annotate the anchors improves the performance of the main method by about $2-3\%$. On the other hand, the DEXTR segmentation is significantly faster in practice, and yields satisfactory results.

4.2.5 Qualitative analysis: We further perform qualitative analysis of segmentation mask selection protocol. Figure 6 shows the effectiveness of our proposed mask selection strategy on a challenging sequence from the VOT dataset [11]. Every 10th frame of the video is manually annotated and the segmentation for the remaining frames is determined by a forward and backward propagation. The proposed selection method almost always selects the better mask. In cases where the selection fails, both masks are similarly good or bad. Examples of these situations are visualized at the second and third arrow in Figure 6, where the J&F score of selected mask is worse, but segmentation masks are still equally good. In contrast, at last arrow, where higher scoring segmentation mask is selected, it is clearly better.

A qualitative analysis on three sequences is performed for additional insights : (i) the first sequence depicts tracking a rectangular *book* with out-of-plane rotations and folding, (ii) the second sequence depicts a *car* with substantial blurring, and (iii) the third sequence depicts an articulated body of an *ice skater*. The sequences are approximately 150 frames long and every 50th frame is annotated by the user, by inputting either 4 extreme points or scribbles in case of IVS [18] and IVOS [10].

Results given in Figure 9, show that IVS [18] performs worst in terms of robustness, *i.e.*, it completely looses track in case of the *book*. IVOS [10] also struggles as it begins to drift to the hands of the girl holding the book. gVOST does not lose track of the object, nor it segments surrounding regions belonging to hands as part of the object. A similar phenomenon is observed for the *car* sequence, where IVS starts to lose the object frame by frame. With the ice skater sequence IVOS struggles with the space between the legs of the iceskater and labels it as part of the body. IVS performs a better, but still includes some background as part of the person. gVOST segments all three objects most accurately with a minimal user interaction.

4.3 Evaluation on the DAVIS2017 dataset

The video object segmentation datasets typically contain shorter sequences with large objects which do not visually change as significantly as in the tracking sequences. The evaluation is more focused on accuracy. To demonstrate the generality of gVOST, it is further evaluated on the reference dataset for interactive video object segmentation [20].

4.3.1 Experimental Setup: The original DAVIS interactive experiment protocol [20] is used. It allows a limited set of user interaction steps for segmenting the entire video. The first interaction step

involves annotating the first frame of the sequence. The method is then run to propagate the annotated mask to all the remaining frames. In each subsequent interaction step, the frame with the worst segmentation mask is selected for re-annotation and the masks are propagated again. The annotation experiment is stopped after eight interactions. An additional experiment using the protocol presented in Subsection 4.2 is also made.

4.3.2 Results: Table 3 compares the overall results of the gVOST in comparison to state-of-the-art. gVOST outperforms the winner of the DAVIS2018 interactive challenge IVS [18] as well as other methods, and performs similarly to the winner of the DAVIS2020 interactive challenge [10].

Method	IoU	J&F
Najafi *et al.* [17]	0.548	-
Heo *et al.* [9]	0.725	0.752
IVS [18]	0.734	-
IVOS [10]	0.790	0.827
gVOST	0.745	0.775

Table 3. Video object segmentation results on the DAVIS 2017 validation dataset after 8 initializations. gVOST outperforms the winner of the DAVIS 2018 interactive challenge in IoU.

Results for the interaction protocol given in Section 4.3.1 are shown in Figure 7. gVOST performs on par with the IVS [18] in terms of IoU, and it outperforms it in terms of J&F. gVOST$_{\mathrm{GT}}$ reaches 0.83 IoU at just 5 fully annotated objects on sequence, meanwhile the winner of DAVIS 2020 interactive challenge IVOS [10] reaches the same performance at 40 scribble interactions per object on sequence. We argue that annotating with gVOST$_{\mathrm{GT}}$ is more efficient. Though gVOST$_{\mathrm{GT}}$ takes six minutes of the user labour (if segmentation masks are created fully manually, not in combination with DEXTR), and IVOS needs approximately five minutes of user labour, searching the worst annotated frame in sequence, interaction and propagation procedure with IVOS has to be repeated 40 times. With gVOST$_{\mathrm{GT}}$ the procedure is repeated only five times, resulting in easier and more efficient sequence annotation.

5 Conclusion

A novel method for user-guided video object segmentation, gVOST is presented. In the first stage a sparse subset of anchor frames are selected and segmented by a user using an interactive image segmentation technique; then anchor labels are propagated from the anchors to the rest of the frames using a visual object tracking algorithm and a proposal selection mechanism. To the best of knowledge, gVOST is the first method designed for annotation in challenging videos.

The effectiveness of our approach is demonstrated on a visual object tracking dataset VOT2020 [11], achieveing 0.73 IoU at just 5% of annotated frames on the dataset, gVOST outperforms the state-of-the-art interactive video object segmentation methods IVS [18] and IVOS [10] by 67% and 40%, respectively. By annotating approximately 10% of the frames, our method achieves the accuracy level, comparable to the quality of ground truth. It can be concluded that the the presented method outperforms the current interactive video object segmentation methods, which have worse tracking ability and tend to lose track of the object in challenging sequences that are common in realistic real world scenarios. On the reference interactive video object segmentation datasets [20] gVOST performs comparably to state-of-the-art, thus demonstrating generality over a range of video types.

A video segmentation application is developed, which implements the gVOST method. The application offers an intuitive graphical user interface and can be used without extensive computer vision experience. Our application significantly reduces the user involvement in annotation of video sequences. The time taken to annotate a single video sequence can be on average lowered by 98% in comparison with manual annotation. In other words, more than six hours used for manually segmenting out an average tracking video sequence reduces down to approximately 5 minutes of user labour.

The application has already been used by independant researchers to annotate of a the VOT2021 tracking dataset. They find it very useful and time efficient. Based on the detailed feedback, we have identified potential improvements of the method will be investigated in future work. For example, after an annotation iteration, the user reviews the whole sequence to find and correct segments by adding new anchors. This could be improved by developing an algorithm that would automatically suggest which frames are very likely to have been poorly segmented. The image segmentation method will be modified to allow manual correction with *negative* clicks, enabling manual exclusion of wrongly segmented parts.

Acknowledgements

The work was supported in part by the Slovenian Re-

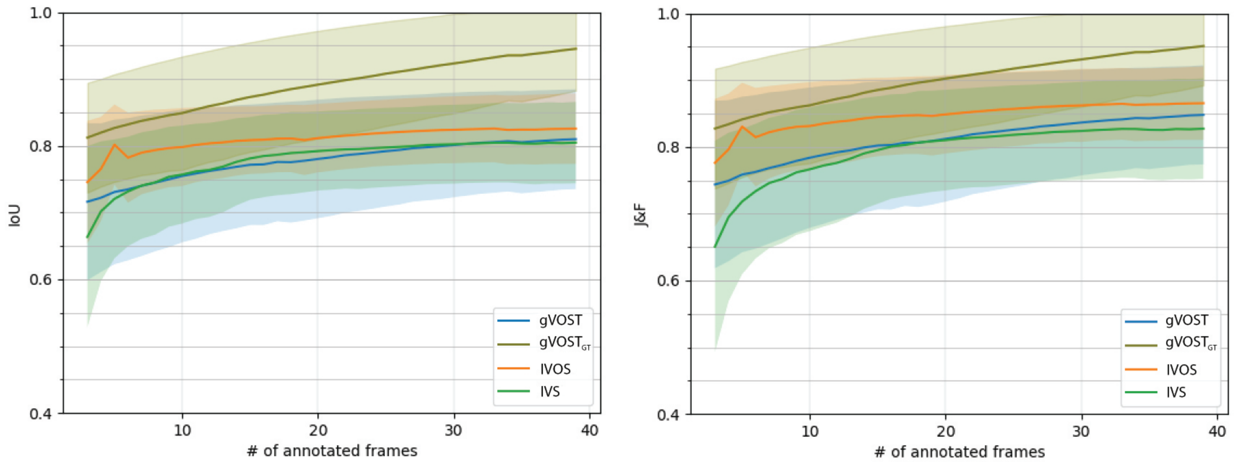


Figure 7. IoU and J&F performances on the validation set of the DAVIS2017 dataset according to number of annotated frames.

search Agency (ARRS), research projects J2-2506 and Z2-1866, and the ARRS research program P2-0214.

References

[1] Sergi Caelles et al. “One-Shot Video Object Segmentation”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2017.

[2] Sergi Caelles et al. “The 2018 DAVIS Challenge on Video Object Segmentation”. In: *CoRR* abs/1803.00557 (2018). arXiv: 1803.00557. URL: http://arxiv.org/abs/1803.00557.

[3] Bowen Chen et al. “ScribbleBox: Interactive Annotation Framework for Video Object Segmentation”. In: *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII*. Ed. by Andrea Vedaldi et al. Vol. 12358. Lecture Notes in Computer Science. Springer, 2020, pp. 293–310. DOI: 10.1007/978-3-030-58601-0_18. URL: https://doi.org/10.1007/978-3-030-58601-0%5C_18.

[4] M. Danelljan et al. “ATOM: Accurate Tracking by Overlap Maximization”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4655–4664. DOI: 10.1109/CVPR.2019.00479.

[5] Mark Everingham et al. “The Pascal Visual Object Classes Challenge: A Retrospective”. In: *Int. J. Comput. Vision* 111.1 (Jan. 2015), 98â€“136. ISSN: 0920-5691. DOI: 10.1007/s11263-014-0733-5. URL: https://doi.org/10.1007/s11263-014-0733-5.

[6] Monica Gruosso, Nicola Capece, and Ugo Erra. “Human segmentation in surveillance video with deep learning”. In: *Multimedia Tools and Applications* 80.1 (2021), pp. 1175–1199.

[7] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90. URL: https://doi.org/10.1109/CVPR.2016.90.

[8] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

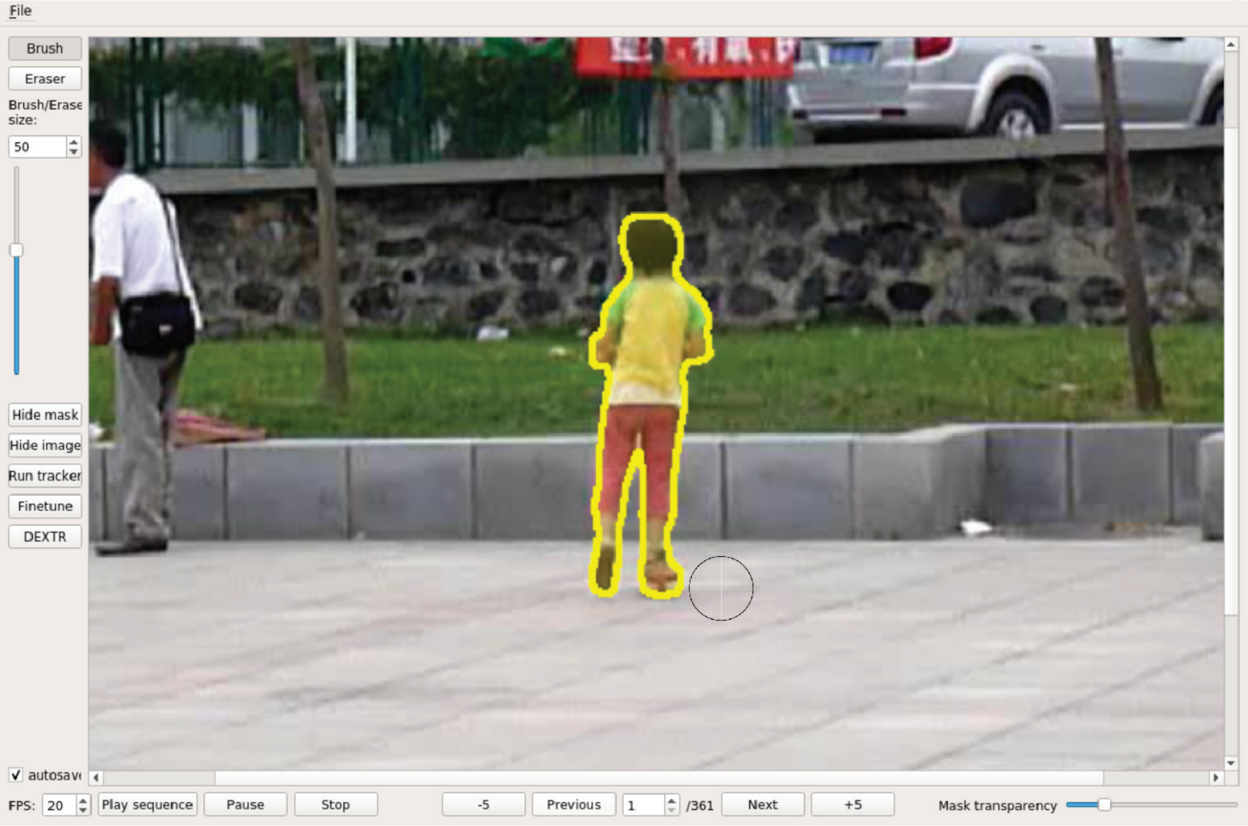
[9] Y. Heo, Y. J. Koh, and C. Kim. “Interactive Video Object Segmentation Using Sparse-to-Dense Networks”. In: *The 2019 DAVIS Challenge on Video Object Segmentation - CVPR Workshops* (2019).

[10] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. “Interactive Video Object Segmentation Using Global and Local Transfer Modules”. In: *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*. Ed. by Andrea Vedaldi et al. Vol. 12362. Lecture Notes in Computer Science. Springer, 2020, pp. 297–313. DOI: 10.1007/978-3-030-58520-4_18. URL: https://doi.org/10.1007/978-3-030-58520-4%5C_18.

[11] Matej Kristan et al. *The Eighth Visual Object Tracking VOT2020 Challenge Results*. 2020.

[12] Di Lin et al. “ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 3159–3167. DOI: 10.1109/CVPR.2016.344. URL: https://doi.org/10.1109/CVPR.2016.344.

[13] Zheng Lin et al. “Interactive Image Segmentation With First Click Attention”. In: *Proceedings of*



Brush/Erase size: 50

☑ autosav

FPS: 20

1 /361

Figure 8. The graphical user interface of the application. It is composed of the file menu bar on the top, and the toolbar that is located on left and bottom border.

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2020.

[14] Alan Lukezic, Jiri Matas, and Matej Kristan. "D3S - A Discriminative Single Shot Segmentation Tracker". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 7131–7140. DOI: 10.1109/CVPR42600.2020.00716. URL: https://doi.org/10.1109/CVPR42600.2020.00716.

[15] K.-K. Maninis et al. "Deep Extreme Cut: From Extreme Points to Object Segmentation". In: *Computer Vision and Pattern Recognition (CVPR)*. 2018.

[16] Eslam Mohamed et al. "Monocular Instance Motion Segmentation for Autonomous Driving: KITTI InstanceMotSeg Dataset and Multi-task Baseline". In: *arXiv preprint arXiv:2008.07008* (2020).

[17] M. Najafi, V. Kulharia T. Ajanthan, and P. H. S. Torr. "Similarity Learning for Dense Label Transfer". In: *The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops* (2018).

[18] Seoung Wug Oh et al. "Fast User-Guided Video Object Segmentation by Interaction-And-Propagation Networks". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 5247–5256. DOI: 10.1109/CVPR.2019.00539.

[19] F. Perazzi et al. "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 724–732. DOI: 10.1109/CVPR.2016.85.

[20] Jordi Pont-Tuset et al. "The 2017 DAVIS Challenge on Video Object Segmentation". In: *CoRR* abs/1704.00675 (2017). arXiv: 1704.00675. URL: http://arxiv.org/abs/1704.00675.

[21] H. Ren, Y. Yang, and X. Liu. "Robust Multiple Object Mask Propagation with Efficient Object Tracking". In: *The 2019 DAVIS Challenge on Video Object Segmentation - CVPR Workshops* (2019).

[22] A. Shaban et al. "Multiple-Instance Video Segmentation with Sequence-Specific Object Proposals". In: *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops* (2017).

[23] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale

Image Recognition". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1409.1556.

[24] Yanan Sun et al. "Deep Video Matting via Spatio-Temporal Alignment and Aggregation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 6975–6984.

[25] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. "Learning Video Object Segmentation With Visual Memory". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017.

[26] P. Voigtlaender and B. Leibe. "Online Adaptation of Convolutional Neural Networks for the 2017 DAVIS Challenge on Video Object Segmentation". In: *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops* (2017).

[27] J. Wu et al. "MILCut: A Sweeping Line Multiple Instance Learning Paradigm for Interactive Image Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 256–263. DOI: 10.1109/CVPR.2014.40.

[28] Ning Xu et al. "YouTube-VOS: Sequence-to-Sequence Video Object Segmentation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018.

[29] S. Zhang et al. "Interactive Object Segmentation With Inside-Outside Guidance". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 12231–12241. DOI: 10.1109/CVPR42600.2020.01225.

Jer Pelhan received his masters degree from the Faculty of Computer and Information Science, University of Ljubljana, Slovenia, in 2023. He is currently a researcher at the same faculty. His research interests include few-shot learning, object detection and segmentation, object counting and object visual tracking.

Matej Kristan received his Ph.D from the Faculty of Electrical Engineering, University of Ljubljana, in 2008. He is a full professor and a vice chair of the department of artificial intelligence at the Faculty of Computer and Information Science. He is president of the IAPR Slovenian pattern recognition society and an Associate Editor of IJCV. His research interests include visual object tracking, anomaly detection, object detection and segmentation, perception methods for autonomous boats and physics-informed machine-learning.

Alan Lukežič received his Ph.D. degree from the Faculty of Computer and Information Science, University of Ljubljana, Slovenia in 2021. He is currently with the Visual Cognitive Systems Laboratory, Faculty of Computer and Information Science, University of Ljubljana, as a Teaching Assistant and a Researcher. His research interests include computer vision, data mining and machine learning.

Jiri Matas is a professor at the Center for Machine Perception, Czech Technical University in Prague. He holds a PhD degree from the University of Surrey, UK (1995). He has published over 200 scientific papers. His publications have approximately 34000 citations in Google Scholar and 13000 in the Web of Science. His h-index is 65 (Google scholar) and 43 (Clarivate Analytics Web of Science) respectively. He received the best paper prize at the BMVC in 2002 and 2005, at the ACCV in 2007 and at ICDAR in 2015. He is on the editorial board of IJCV and was an Associate Editor-in-Chief of IEEE TPAMI. His research interests include visual tracking, object recognition, image matching and retrieval, sequential pattern recognition, and RANSAC-type optimization methods.

Luka Čehovin Zajc received his Ph.D degree from the Faculty of Computer and Information Science, University of Ljubljana, Slovenia in 2015. He is working at the Visual Cognitive Systems Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Slovenia as an assistant professor and a researcher. His research interests include computer vision, educational robotics, human-computer interaction, and spatial computing.

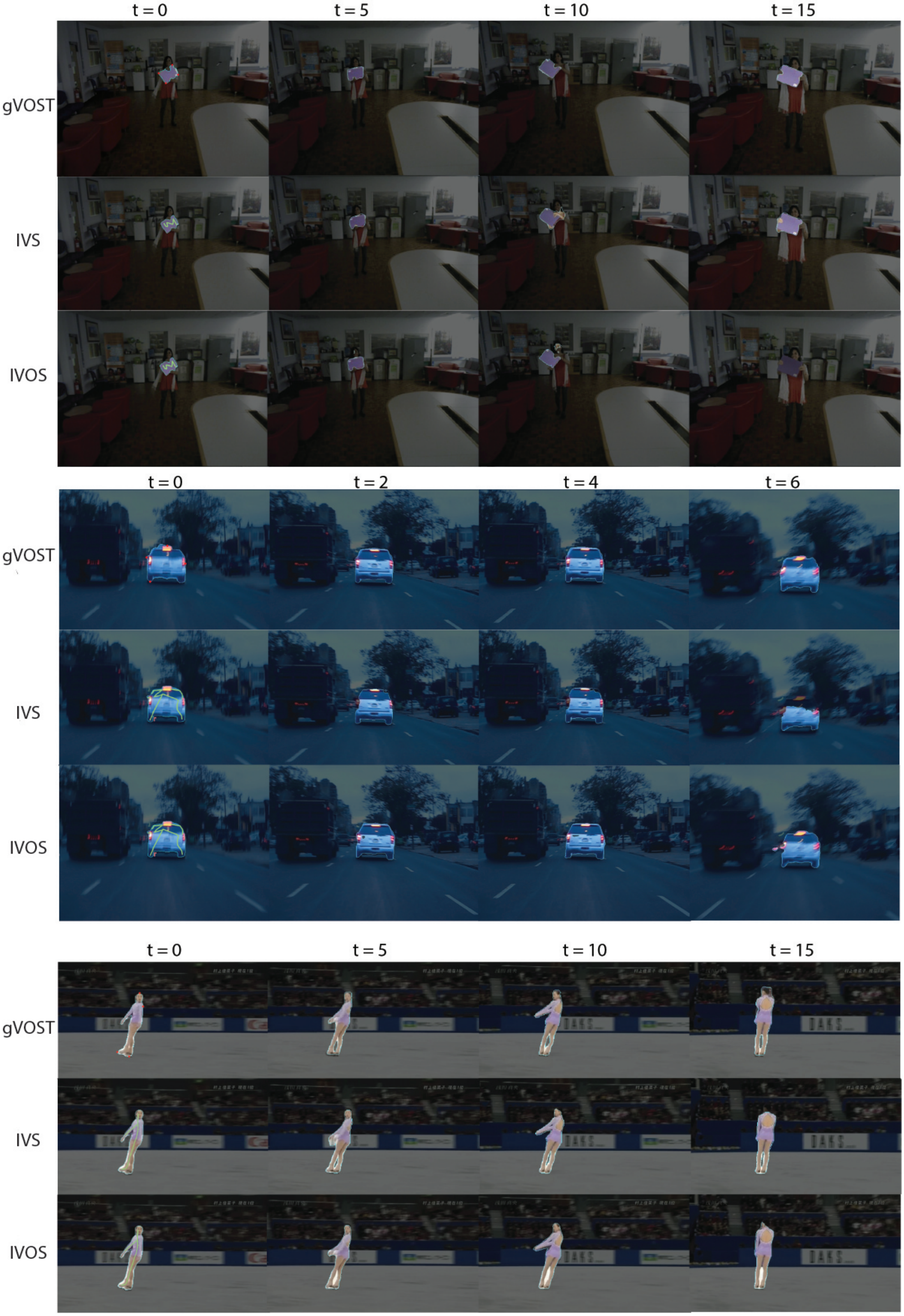


Figure 9. Selected results on the VOT2020 dataset. First column presents initialization frames.