Noname manuscript No.
(will be inserted by the editor)

Contextual Guided Segmentation Framework for Semi-supervised Video Instance Segmentation

Trung-Nghia Le · Tam V. Nguyen · Minh-Triet Tran

Received: date / Accepted: date

Abstract In this paper, we propose Contextual Guided Segmentation (CGS) framework for video instance segmentation in three passes. In the first pass, *i.e*. preview segmentation, we propose Instance Re-Identification Flow to estimate main properties of each instance (i.e., human/non-human, rigid/deformable, known/unknown category) by propagating its preview mask to other frames. In the second pass, *i.e*. contextual segmentation, we introduce multiple contextual segmentation schemes. For human instance, we develop skeleton-guided segmentation in a frame along with object flow to correct and refine the result across frames. For non-human instance, if the instance has a wide variation in appearance and belongs to known categories (which can be inferred from the initial mask), we adopt instance segmentation. If the non-human instance is nearly rigid, we train FCNs on synthesized images from the first frame of a video sequence. In the final pass, *i.e*. guided segmentation, we develop a novel fined-grained segmentation method on non-rectangular regions of interest (ROIs). The natural-shaped ROI is generated by applying guided attention from the neighbor frames of the current one to reduce the ambiguity in the segmentation of different overlapping instances. Forward mask propagation is followed by backward mask propagation to further restore missing instance fragments due to reappeared instances, fast motion, occlusion, or heavy deformation. Finally, instances in each frame are merged based on their depth values, together with human and non-human object interaction and rare instance priority. Experiments conducted on the DAVIS Test-Challenge dataset demonstrate the effectiveness of our proposed framework. We achieved the 3^{rd} consistently in the DAVIS Challenges 2017-2019 with

Trung-Nghia Le
National Institute of Informatics, Tokyo, Japan. E-mail: ltnghia@nii.ac.jp

Tam V. Nguyen
Department of Computer Science, University of Dayton, Ohio, USA. E-mail: tamnguyen@udayton.edu

Minh-Triet Tran
Corresponding author, University of Science and Vietnam National University, Ho Chi Minh, Vietnam. E-mail: tmtriet@fit.hcmus.edu.vn

Fig. 1 Examples of results obtained by our proposed method. From left to right: the first video frame with the ground-truth label followed by results of our method on next frames.

75.4%, 72.4%, and 78.4% in terms of global score, region similarity, and contour accuracy, respectively.

Keywords Semi-supervised learning · Video object segmentation · Contextual segmentation · Guided segmentation.

1 Introduction

Object segmentation is considered a labeling problem aiming to separate foreground from background regions. Video instance segmentation, which is higher-level and more challenging than object segmentation, aims to label each video frame pixel to instances or the background region and then assign consistent IDs to these instances over the video sequence. Object/instance segmentation in videos is beneficial in a wide range of practical applications, *i.e.*, autonomous vehicle [1], action recognition [21], video summarization [30], object tracking [66], scene understanding [70], and video annotation [28].

This paper focuses on semi-supervised video instance segmentation [46], which targets certain instances whose ground-truth mask for the first video frame is given. DAVIS Challenge [46] promotes the development of this task. The benchmark dataset of this challenge consists of many pitfalls such as rapid motion, distractors, smaller objects, fine structures, occlusions, large deformations, complex object interactions, and so on. Figure 1 shows some exemplary results of our proposed method on the DAVIS Test-Challenge dataset [46].

To address the challenges of the given problem, tracking and re-identification methods are adopted and jointly integrated into segmentation models to keep the consistency of targeted instances over the entire video sequence [17,22,31,32]. However, existing works usually fail to follow and segment targeted instances due to cannot cover all various contexts in the video. We argue that context information is essential for semantic segmentation to reduce ambiguous instances and obtain robust results. Therefore, this work aims to leverage the context information to improve the performance of video instance segmentation. Inspired by the idea of *"you should look*

twice" [42, 43] in the task of object detection, we propose a three-pass guided segmentation framework, namely Contextual Guided Segmentation (CGS), to tackle the problem of semi-supervised video instance segmentation. Our proposed method consists of two key ideas as below.

First, we exploit variation in the video and propose various contextual segmentation strategies adapting to contexts, *i.e*. the category and visual properties of an instance. To select the appropriate scheme, we propose a novel Instance Re-Identification Flow (IRIF) to propagate the initial mask of an instance to other frames and analyze the visual properties of segmented regions. Multiple contextual segmentation schemes are also introduced to adapt the contextual properties of each instance. For human instances, we develop skeleton-guided segmentation. For non-human instances, we train FCNs from our synthesized dataset for nearly-rigid instances with similar background scenes. Instance segmentation detectors are utilized to handle deformable non-human instances in known categories. Results from our IRIF are treated as the baseline scheme for other cases.

Second, to segment an instance in a region of interest (ROI), we propose novel guided fined-grained segmentation based on attention for performance improvement. We transform a regular rectangular ROI to a non-rectangular ROI by blending attention inferred from neighbor frames to eliminate complex background inside the ROI. We also propose bi-directional propagation strategies to construct adaptive attention for guided segmentation. Forward propagation strategy can correct missing segmentation due to dense objects in a ROI. Meanwhile, a backward propagation strategy can recover missing instances due to fast motion, occlusion, or heavy deformation.

The DAVIS Challenges 2017-2019 results indicate that our method is competitive among the top-performing submissions. Our early results were preliminarily listed on DAVIS 2017 Challenge [25], DAVIS 2018 Challenge [59], and DAVIS 2019 Challenge [58]. In this paper, we provide the full details of our proposed framework. Our contributions are as follows.

- We propose Contextual Guided Segmentation (CGS) framework with three segmentation passes to exploit various contexts in video instance segmentation. Our proposed method achieved the $3rd^{th}$ ranking consistently in the DAVIS Challenges 2017-2019.
- We propose Instance Re-Identification Flow (IRIF) to extract contextual properties of each instance by propagating its preview mask from the current frame to coming frames.
- We introduce multiple contextual segmentation schemes to adapt the contextual properties of each instance.
- We propose bi-directional propagation strategies for guided fined-grained segmentation in non-rectangular ROIs. Our proposed guided segmentation outperforms the standard segmentation, which is mostly applied in rectangular ROIs.
- To blend instance masks into a unique result, we introduce a merging process based on their depth values together with human and non-human object interaction and rare instance priority.
- We construct Wonderland Data to increase the number of training data for one-shot learning. Our proposed augmentation approach also can be utilized for different problems.

The remainder of this paper is organized as follows. In Section 2, we briefly review the related work. Next, our proposed methods are presented in Section 3. Experimental results are then reported and discussed in Section 4. Finally, Section 5 concludes and paves the way for future work.

2 Related Work

2.1 One-Shot Learning

Data augmentation is essential to deal with one-shot learning [2], which aims to train a deep network with only a given first video frame. Caelles *et al*. [2] introduced the first simple data augmentation strategy such as random crop, random scale, vertical flip, random changes in brightness, saturation, and contrast of the given first frame. Khoreva *et al*. [22] later introduce Lucid Dreaming [22] to synthesize the foreground changes by rigid and non-rigid transformation with a small extent, and synthesize the background changes using affine deformations with limited appearance variations. The given first frame with ground truth is augmented with Lucid Dreaming to generate more training data with different viewpoints, leading to much improvement of training networks. Hence, augmented data by Lucid Dreaming, called Lucid Data, has become common for one-shot learning. However, Lucid Data cannot deal with different backgrounds caused by objects' motion or camera view changes. Guo *et al*. [17] changed the background of the first video frame by images with pure background crawled randomly from the Internet by Google, namely Online Data. However, Online Data is unstable because of randomly crawled from the Internet without considering the content of the video. Meanwhile, our Wonderland Data is filtered out from large-scale scene data to choose the most similar scenes with the video.

Khoreva *et al*. [22] trained appearance-based and motion-based models with Lucid Data [22]. Shaban *et al*. [54] learned video segments by bootstrapping them from temporally consistent object proposals, which are first spatially trained on Lucid Data [22] and then incorporated a semi-Markov pixel-level motion model to form spatio-temporal object proposals. Luiten *et al*. [38] first trained DeepLab3+ [8] on a combination of standard datasets and then fine-tuned the network on Lucid Data [22] of each video to form a strong network to segment instance inside ROI. Li *et al*. [31] trained online re-identification network, which is the original Region Proposal Network of Mask R-CNN, and a recurrent mask propagation network on Lucid Data [22]. Xu [71] proposed a spatio-temporal CNN in which the spatial segmentation branch is fine-tuned online on Lucid Data of each sequence while the temporal coherence branch is trained offline on the entire dataset. Models are not only fine-tuned offline on Lucid Data [22] of the first frame but also can be updated online while processing the video [62]. Mask R-CNN is fine-tuned on Lucid Data [74] or Online Data [17] to adapt proposals to the video.

2.2 Temporal Connection Mining

This approach aims to perform instance tracking, propagation, and re-identification, where each instance is detected and re-identified through frames [32]. Li *et al*. [32] iteratively propagated masks via flow warping and re-identified instances via adaptive matching to retrieve missing ones. Luiten *et al*. [38] first segmented multiple object proposals in the entire video and then selected and linked these proposals over time using a re-identification feature embedding vector for each proposal. Re-identification feature embedding vectors are computed using a triplet-loss based re-identification embedding network. Li *et al*. [31] jointed re-identification and attention-based recurrent temporal propagation into a unified framework to retrieve missing objects despite their large appearance changes. Guo *et al*. [17] first extracted possible mask proposals in each frame and then joined tracking and re-identification to filter and rank proposals to merge the highest confident proposals. Xu *et al*. [74] adapted a multiple hypotheses tracking method to build up a bounding box proposal tracking tree for different objects, then propagate masks, and finally merged mask proposals from the tracking tree. Wang *et al*. [66] used fully convolutional Siamese trackers to produce class-agnostic binary segmentation masks of the target objects. Voigtlaender *et al*. [61] used a semantic pixel-wise embedding together with a global and a local matching mechanism to transfer information from the first frame and from the previous frame of the video to the current frame, which is used as internal guidance for segmentation. Jonathon *et al*. [39] used a Siamese architecture to detect and track multiple objects and then performed segmentation inside the detected bounding boxes. Tran *et al*. [57] propagated masks with reference to multiple extra samples through a memory reference pool.

2.3 End-to-End Temporal Learning

This approach directly learns temporal information in a video through deep learning architectures such as LSTM, guided-attention, or memory networks. Some methods combine feature maps from different video frames by correlation matching [61] or non-local matching [44]. Guo *et al*. [16] integrated STM [44] into DeepLabv3+ [8] to concatenate low-level features in mask decoder. Andreas *et al*. [48] implemented a memory network to add semantic information about the target object from a previous frame to the refinement stage, complementing the predictions provided by the target appearance model. Zhang *et al*. [78] developed a spatial constraint module that takes the previous prediction to generate a spatial prior for the current frame, helping to disambiguate appearance confusion and eliminate false predictions. Fiaz *et al*. [15] introduced a guided feature learning without model update algorithm for directional deep appearance learning. Liu *et al*. [35] integrated multilevel backbone into memory network to generate higher spatial resolution features. Le *et al*. [64] leveraged existing memory-based models and enhanced their capability by adding pre-processing and post-processing steps. Xie *et al*. [69] integrated depth maps from a video sequence into STM [44] to alleviate the ambiguity of objects with similar appearances. Seong *et al*. [53] developed a kernelized memory network and used the Hide-and-Seek strat-

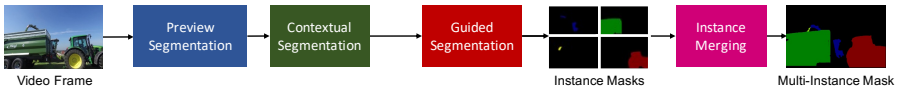


Fig. 2 Overview of our Contextual Guided Segmentation (CGS) framework.

egy training to handle occlusions and segment boundary extraction. Yang *et al*. [77] combined collaborative foreground-background integration with multi-scale matching to be robust to various object scales.

3 Proposed Method

3.1 **Overview**

Figure 2 illustrates CGS with three passes: preview segmentation for context evaluation, contextual segmentation, guided segmentation based on propagation. In particular, In the first pass, we propose Instance Re-Identification Flow (IRIF) to generate the preview mask sequence and extract different contextual properties from each instance. In the second pass, we introduce multiple segmentation schemes corresponding to extracted properties. In the third pass, we develop fined-grained segmentation based on guided propagation. We remark that each instance is processed independently over frames of a video sequence. Finally, instance masks are then blended with reference to depth information, human and non-human instance interaction, and rare instance priority.

3.2 **Preview Segmentation**

Figure 3 illustrates the flow chart of Instance Re-Identification Flow (IRIF) for preview segmentation. The segmentation performed on the current frame is based on the history information of the previous frames. The segmentation result of the current frame is further fed to the process of the coming frame.

We remark that in this component, we consider two types of instance, *i.e*. human and non-human, to treat each instance in different ways. Given the first frame with its ground truth label, we extract the bounding box for each instance and then perform human/non-human classification for all instances using Mask R-CNN [18].

3.2.1 Instance Localization and Tracking

For each video frame, we localize and track instances in a re-identification manner. Note that we expand the bounding box to 10% to well capture the whole area of the object instances. For *human objects*, we employ person search [68] by detecting person by using Faster R-CNN and then extracting person re-identification feature for all detected person region. On the other hand, DeepFlow [67] and Deformable Part Models (DPM) [14] are utilized to detect and track *non-human objects*.

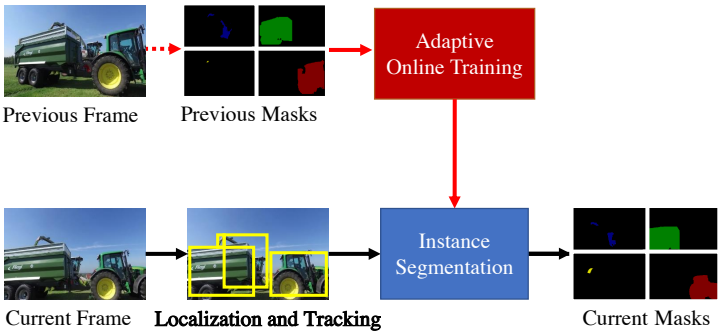


Fig. 3 The flowchart of Instance Re-Identification Flow (IRIF) component. The segmentation performed on the current frame is based on the history information of the previous frames. The segmentation result of the current frame is further fed to the process of the coming frame.

3.2.2 Adaptive Online Learning for Instance Segmentation

For each instance, to identify each pixel as foreground (instance) or background, we utilize multiple binary SVM classifiers [6] which is learned from the appearance of the previous n frames with sampling step size δ, where n and δ are set as 8 and 2, respectively. Note that our multiple binary SVM classifiers are implemented for history reference with several unary instances, *e.g.*, saliency [36], CNN features [23], location of the bounding box, and color, to segment each instance within its tracked bounding box in each frame. We only update the SVM model if the size of one instance significantly changes. We then utilize GrabCut [49] for each instance to separate it from the background. After this step, each pixel is assigned with the instance ID.

Specifically for human instance, in case the instance is missing and re-appears in the next couple of frames, we adopt the state-of-the-art image parser, Pyramid Scene Parsing (PSPNet) [81] with the pre-trained model on PASCAL VOC dataset [13]. The re-identification results from PSPNet are blended into our segmentation outcomes.

3.2.3 Contextual Property Extraction

This component aims to determine the context of an instance so that we can apply an appropriate segmentation scheme for that instance. The context can be any observable properties that may affect the strategy to extract the mask of an instance in frames efficiently. In this work, we consider the following three attributes of an instance as its context: human or non-human, known or unknown category, rigid or deformable.

The category of an instance, such as person, car, dog, etc., can be directly inferred from its initial mask using pre-trained Mask R-CNN [18] on the MS-COCO dataset.

To evaluate if an instance is rigid or deformable, we analyze the preview sequence of instance masks in the first $n_{Preview}$ frames. If there exists a homography matrix to transform the instance from the first frame to another frame for most frames in the first $n_{Preview}$ frames, we consider the instance to be rigid.

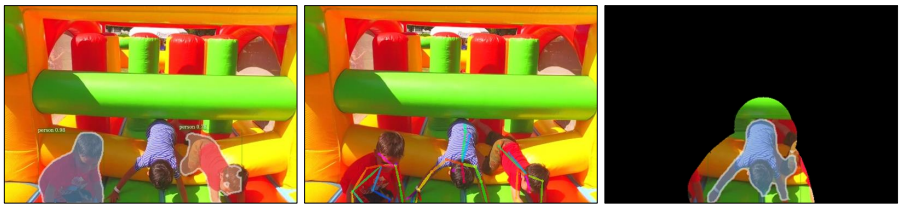


Fig. 4 Skeleton-guided segmentation for unusual pose.

3.3 **Contextual Segmentation**

Each instance is segmented in different appropriate ways in this contextual segmentation, adapting to its extracted contextual properties (i.e., human/non-human, rigid/deformable, known/unknown category).

3.3.1 Human Instance Segmentation

We employ Mask R-CNN [18], pre-trained on the MS-COCO dataset [34], to extract human segments. However, the results of Mask R-CNN may be affected by occlusion or unusual human pose.

To overcome this issue, we develop *skeleton-guided segmentation*. We use the skeletons from OpenPose [4] for reference to control and refine human instance segmentation. For a human instance with an unusual pose that Mask R-CNN cannot recognize, we dilate the skeleton to obtain a skeleton-guided region, *i.e*. an image with only the region containing the complete human instance. We then apply Mask R-CNN on a skeleton-guided region. By eliminating unrelated content, Mask R-CNN has a higher chance to extract human instance segment correctly (see Fig. 4). To preserve the inter-frame mask consistency, we use object flow [60] to correct and refine the result across frames.

3.3.2 Rigid Non-Human Instance Segmentation

For this type of instance, our objective is to accurately extract such instances from different backgrounds in the same scene category with the initial frame. Our method to process each instance is as follows. First, we synthesize images from the first frame of a video sequence, resulting in Wonderland Data. Second, to segment instances inside bounding boxes, we train DeepLab2 [7] and OSVOS [2] on our synthesized Wonderland Data.

Wonderland Data Generation: Differently from existing work, we exploit various contextual properties from instances. After that, multiple segmentation schemes are performed for each instance, adapting to its extracted contextual properties. Inspire by Lucid Data [22], we introduce new augmented data, namely Wonderland

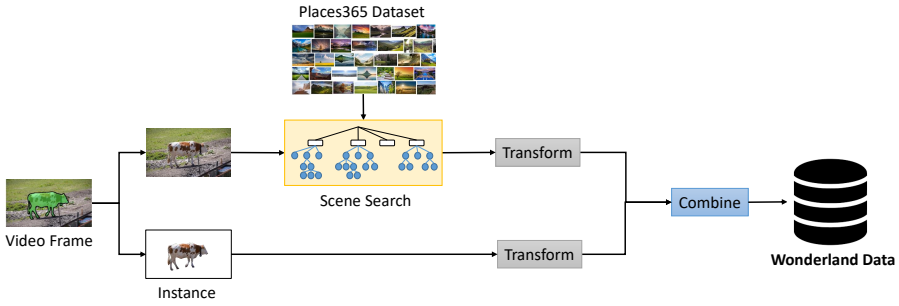


Fig. 5 Wonderland Data generation.

Fig. 6 Augmented data generated by different methods. From left to right: the original video frames with overlaid ground-truth, followed by corresponding Lucid Data [22] and our proposed Wonderland Data in this order.

Data. To generate visual variations of the initial mask, we apply both affine and non-rigid deformations, together with illumination changes, on the mask. We also replace the background with most similar scenes filtered out from a large-scale Places365 dataset [82] to preserve the semantics of the image. In this way, we can increase more training samples than Lucid Data (10,000 images for each video, in comparing with 2,500 images of Lucid Data) to deal with one-shot learning.

Figure 5 illustrates our proposed Wonderland Data generation. In this work, from a pair of an input image and a mask, we generate $10,000$ different pairs of synthesized images and masks. The Wonderland Data is published on our website[1]. We collect scene photos from the training set of the Places365 dataset [82], which has about 8 million images divided into 365 scene categories. We manually discard artificial scenes, use only 22 natural scene categories with 592k images. For each image, we extract a feature at the last layer of DenseNet-161 [20], which was pre-trained on the Places365 dataset [82]. This feature is used to build a hierarchical k-mean search for

[1] https://sites.google.com/view/ltnghia/research/vos

Fig. 7 The flowchart of our network training process.

each category independently. We assume that each node has M images, and a leaf node has maximum L images. To cluster images at a node, we propose to use K-mean algorithm with $K = \min(M \backslash L, T)$. In this work, we empirically set $L = 200$ and $T = 200$ to speed up clustering.

We classify an input image into the corresponding category, using the pre-trained DenseNet-161 on the Places365 challenge dataset. We also extract a channel feature at the last layer of the same network. After that, we search leaf nodes by comparing the Euclidean distance between the feature of an input image and the center of clusters. To search N images, we randomly choose 80% number of images of the nearest leaf node and $70\%, 60\%, 50\%$, etc. number of images of next leaf nodes, respectively.

We also extract the object mask from the input image, then transform the object and searched scenes independently, similarly to [22]. In more detail, we use affine transformation (*e.g.*, translation, rotation, and scale) and non-rigid deformations, together with illumination changes. Figure 6 shows examples of Lucid Data and our Wonderland Data.

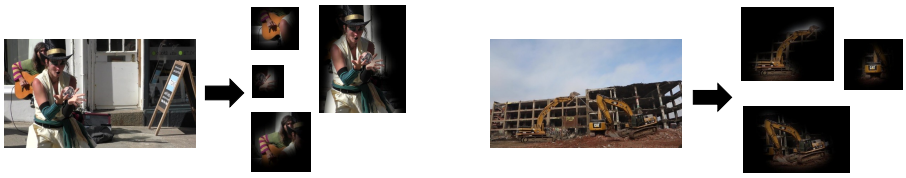
Network Training: Figure 7 our training process, including domain-based training and object-based training. In *domain-based training*, we fine-tune pre-trained networks (*i.e.* DeepLab2 [7] pre-trained on COCO-Stuff dataset [3] and OSVOS [2] pre-trained on ImageNet dataset [50]) on the DAVIS training data for domain transformation. In *object-based training*: we fine-tune networks on the ground-truth mask of each instance of each video. We remark that we use only the first frame of videos and apply the proposed Wonderland Data generation method for these images.

3.3.3 Deformable Non-Human Instance Segmentation

For this instance type, we categorize instances into two groups, namely, known and unknown categories. For the known categories, *i.e.*, already listed in MS-COCO dataset [34], we simply adopt Mask R-CNN to retrieve the instance segments. We directly obtain the preview results from our IRIF component for the unknown categories since it can handle arbitrary object categories.

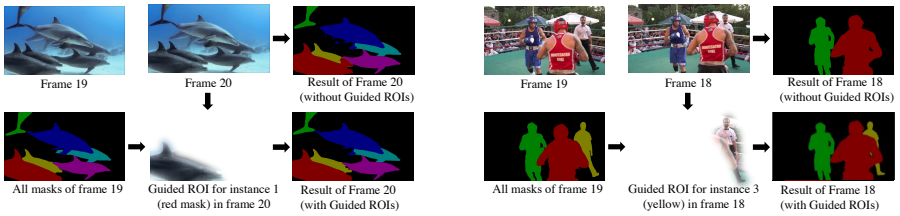
3.4 **Guided Segmentation**

Traditional Fully Convolutional Networks (FCNs) consider the entire rectangular region of interest (ROI) as the input to segment objects inside the ROI. This can lead to incorrect boundary segmentation due to the complex background and concave hull



a) Juggle clip a) Demolition clip

Fig. 8 Visualization of guided non-rectangular ROI.



a) Forward propagation a) Backward propagation

Fig. 9 Visualization of forward and backward propagation.

of the object. To overcome this limitation, we aim to transform a rectangular ROI to a non-rectangular ROI across the object boundary to eliminate the complex background inside the ROI (see Fig. 8). In particular, we utilize referral information from extra frames to identify the shape of the instance of interest inside the ROI of the current frame. We propose to apply guided attention to construct the non-rectangular ROI and then perform fine-grained segmentation on this guided non-rectangular ROI.

3.4.1 Bi-directional Propagation

In particular, we propose bi-directional strategies to construct adaptive attention for guided segmentation. Particularly, initial segments from neighbor frames are used as references for segmentation at the current frame. Attention is computed in two strategies sequentially, *i.e.*, forward propagation and back-propagation, in specific ways adapting the context. Forward propagation strategy, where attention is referenced from initial segments of previous frames, can correct excessed segmentation due to dense objects in a ROI (cf. Fig. 9a). Meanwhile, the back-propagation strategy, where attention is referenced from initial segments of next frames, can recover missing instances due to fast motion, occlusion, or heavy deformation (size changing from tiny to large or vice versa) (cf. Fig. 9b).

3.4.2 Guided Non-Rectangular ROI Construction

To construct a guided non-rectangular ROI, we expand the mask of the interest instance at neighbor frames and then transfer and combine them at the current frame.

This guarantees that the ROI can cover the entire interest instance. We do not apply mask propagation to avoid inaccurate flow warping as well as reducing the complexity of computation. Then, we create a smooth transition region (by applying a blurred mask to remove background) for the guided ROI to avoid a clear border between the ROI and background. It is essential to make the segmentation method focus on the interest instance and avoid inaccurate segmentation due to a clear border. We remark that the range of boundary expansion and transition smooth is computed based on the intensity of movement of the instance. Both propagation strategies are performed adaptively if initial segments of the interest instance at the current frame are much different (in appearance or size) from those at neighbor frames or the instance reappears. On the other hand, we only refine the interest instance at the current frame to save the computational cost.

3.4.3 Fine-grained Segmentation

We use Deep Grabcut [72] and Mask R-CNN [18] for fine-grained segmentation in guided non-rectangular ROIs. Inspired by Luiten et.al. [38], we train DeepLab3+ [8] based on Xception-65 [10] backbone on MS-COCO [34] and Mapillary [40] datasets to enhance the network generalization. For Mask R-CNN, we directly use a pre-trained model on MS-COCO [34] dataset.

3.5 **Refinement and Merging**

Through preliminary results, we observe that the initial segmentation is not smooth enough. Therefore, we refine instance masks to improve segmentation quality, using rare instance attention and boundary snapping.

3.5.1 Rare-Instance Attention Refinement

We further refine the results by considering the rare instances. We observe that rare objects are shrunk due to larger objects. To identify rare object instances, we compute each object instance mask percentage in terms of area (provided in the first frame). Instances with a size smaller than 5% the total size of tracking objects are considered rare ones. We assume that a smaller object tends to be small in the whole video. Next, we recover rare object instances by transferring the results produced by the foreground probability obtained from the binary-SVM classifier on each object instance.

3.5.2 Boundary Snapping Refinement

We also adopt boundary snapping [2] to further refine object shapes. In particular, we extract the saliency [36] and the contour [76] from the video frame. The salient pixels close to the contour are snapped.

3.5.3 Topological Order Estimation for Instance Merging

It is essential to determine the topology relationship (in terms of z-order) between multiple instances to sequentially combine corresponding masks of different instances into the final result. We here merge instances based on human and non-human instance interaction, depth values, and rare instance priority heuristics in this order as follows:

- **Human and non-human instance interaction**: We define interaction heuristics as follow: transportation instances (such as horse, bike, motor, surfboard, and skateboard, etc.) are the farthest from the camera; human instance have the middle distance to the camera; and small non-human instances which can be held, bring, touch, etc. are the nearest from the camera. Interacted small non-human instances are localized at the human hand's position using OpenPose [4].
- **Depth values**: We first estimate pixel-wise depth values of the video frame, using DCNF-FCSP [37], and then take the average value for each instance.
- **Rare instance priority**: We notice that rare instances are always the nearest ones from the camera.

4 Experimental Results

4.1 Dataset Benchmark and Metrics

We participated the DAVIS Challenges 2017-2019, Semi-Supervised Track[2,3,4] and evaluated our methods on the *DAVIS Test-Challenge* dataset. The dataset consists of 150 sequences, totaling $10,459$ annotated frames and 376 instances. There are a total of 30 video sequences for testing, and their ground truth not publicly available. Submissions were made through the CodaLab site of the challenge[5]. This dataset is challenging due to multiple object instances with more distractors, *i.e.*, smaller instances and fine structures, more occlusions, and fast motion.

For the evaluation metrics, per-instance measures are used as described in [45]: Region Jaccard (J) and Boundary F measure (F). The overall measures are computed as the mean between J and F, and both are averaged over all objects.

4.2 Results on DAVIS Challenges 2017-2019

4.2.1 DAVIS 2017 Challenge

Due to the time limit, we submitted the proposed IRIF component in the DAVIS 2017 Challenge and achieved 3^{rd} place out of 22 team submissions in this challenge. As shown in Table 1, our proposed IRIF achieves very promising results in the DAVIS

[2] https://davischallenge.org/challenge2017/index.html
[3] https://davischallenge.org/challenge2018/index.html
[4] https://davischallenge.org/challenge2019/index.html
[5] https://competitions.codalab.org/competitions/21650

Table 1 Top global ranking results in the DAVIS Challenges 2017-2019. The best results are marked in **boldface**. Our results are marked in blue. We note that the teams without references do not have publication.

Rank	Method/Team	Year	Global G	Region J			Boundary F		
			Mean ⇑	Mean ⇑	Recall ⇑	Decay ⇓	Mean ⇑	Recall ⇑	Decay ⇓
1	OSS [65]	2019	**76.7**	72.8	81.5	18.9	**80.7**	87.5	21.3
2	BoLTVOS+ [39]	2019	76.2	**72.9**	**81.7**	16.3	79.4	86.7	19.5
3	CGS [58]	2019	75.4	72.4	**81.7**	**11.0**	78.4	**87.6**	**12.9**
4	STM [44]	2019	75.2	72.6	80.9	21.0	77.7	85.0	24.1
5	PremVOS [38]	2018	74.7	71.0	79.5	19.0	78.4	86.7	20.8
6	DyeNet [31]	2018	73.8	71.9	79.4	19.8	75.8	83.0	20.3
7	Theodoruszq	2019	73.1	70.1	77.3	24.8	76.1	84.0	28.3
8	Panday	2019	71.3	67.7	74.8	24.7	75.0	81.2	27.5
9	DLTA [47]	2019	70.6	68.5	78.1	20.3	72.8	84.2	24.0
10	VS-ReID [32]	2017	69.9	67.9	74.6	25.5	71.9	79.1	24.1
11	CAVOS [73]	2018	69.7	66.9	74.1	23.1	72.5	80.3	25.9
12	ODG [17]	2018	69.5	67.5	77.0	15.0	71.5	82.2	18.5
13	PVOS [16]	2019	69.2	66.0	73.4	28.5	72.3	80.4	31.1
14	LucidTracker [22]	2017	67.8	65.1	72.5	27.7	70.6	79.8	30.2
15	Second Pass [59]	2018	66.3	64.1	75.0	11.7	68.6	80.7	13.5
16	First Pass [26]	2017	63.8	61.5	68.6	17.1	66.2	79.0	17.6
17	SPT [54]	2017	61.5	59.8	71.0	21.9	74.6	74.6	23.7
18	FAVOS [33]	2018	60.6	58.4	65.6	26.2	62.9	71.0	29.7
19	MPN [56]	2018	60.1	57.7	64.9	27.2	62.4	71.7	28.1
20	PALC [63]	2018	58.9	56.7	63.1	30.7	61.1	67.6	33.1
21	OnAVOS [62]	2017	57.7	54.8	60.8	60.5	67.2	67.2	34.7
22	SPN [9]	2017	56.9	54.8	60.7	34.4	59.1	66.7	36.1
23	HE-PSPNet [80]	2017	56.9	53.6	59.5	25.3	60.2	67.9	27.6
24	OSVOS-IOFT [41]	2017	55.8	53.8	60.1	37.7	57.8	62.1	42.9
25	TOP [55]	2017	54.8	51.6	56.3	26.8	57.9	64.8	28.8
26	Froma	2017	53.9	50.9	54.9	32.5	57.1	66.2	33.7

Fig. 10 Visualization results on the DAVIS Test-Challenge dataset. From top to bottom: the first video frame with the ground-truth label followed by results of our proposed methods in preview segmentation [26], contextual segmentation [59], and guided segmentation [58]. The ground-truth of the certain video frame is not publicly available. Our CGS results significantly track and segment the instances of interest as annotated in the first frame.

2017 Challenge, namely, 0.615, 0.662, and 0.638 in terms of region similarity (Jaccard index), contour accuracy (F-measure), and global score, respectively. Our results highly indicate that our method is competitive among the state-of-the-art methods in this dataset. Our method maintains the performance as frames evolve, as seen via the best performance in terms of J decay and F decay among the leading submissions in 2017.

Table 2 The performance of different components in our method on the DAVIS Test-Challenge dataset. PS, CS, and GS stand for preview segmentation, contextual segmentation, and guided segmentation, respectively.

Settings			Global Score ⇑	Region J ⇑	Boundary F ⇑
PS	CS	GS			
✓			63.8	61.5	66.2
✓	✓		66.3	64.1	68.6
✓	✓	✓	**75.4**	**72.4**	**78.4**

4.2.2 DAVIS 2018 Challenge

We also had another submission of CIS framework to the DAVIS 2018 Challenge and achieved 6^{th} place out of 41 team submissions in this challenge. Table 1 shows that our CIS achieves promising results, namely, 64.1%, 68.6%, and 66.3% in terms of region similarity (Jaccard index), contour accuracy (F-measure), and global score, respectively. Our method also maintains the best stable performance in terms of J decay and F decay among the leading submissions in 2018.

4.2.3 DAVIS 2019 Challenge

As shown in Table 1, we obtained very competitive results. Our proposed CGS achieved 0.724, 0.784, and 0.754 in terms of region similarity (J), contour accuracy (F), and global score, respectively. Our method achieved the best performance in Decay and Recall of all metrics consistently. Furthermore, we note that our CGS is in top 3 over 4 teams achieving 0.75 in terms of global score in all three years.

4.2.4 Ablation Study

Table 2 shows the results of our proposed framework with different settings. Our proposed CGS (using all three passes) outperforms using only two passes [59] or a pass [26]. This highlights the significant contribution of the second pass and the third pass, which are the multiple contextual segmentation schemes, and guided instance segmentation, respectively. Particularly, contextual segmentation can improve the performance up to 2.5%. Meanwhile, guided segmentation improves contextual segmentation up to 9.1% in the global score.

Figure 10 visualizes segmentation results. From top row to bottom row, we can observe the first video frame and a triple of processed video frames of our proposed methods in preview segmentation [26], contextual segmentation [59], and guided segmentation [58]. Our final CGS results surpass the performance of others and successfully track and segment the key instances. Our framework can even handle camouflaged instances, small instances, and occluded instances.

5 Conclusion

In this paper, we propose the novel CGS framework for semi-supervised instance segmentation in videos with three segmentation passes. In the first pass, we develop

the novel IRIF for preview instance segmentation and extract contextual information. In the second pass, we introduce multiple contextual segmentation schemes to deal with different instance types, such as human/non-human rigid/non-rigid instances in known/unknown object categories. In the final pass, we propose a novel guided fined-grained segmentation based on attention to eliminate complex background inside the region of interest for performance improvement.

Our proposed methods achieve competitive results among the leading submissions in the DAVIS Challenges consistently, *i.e.* 3^{rd} place, 6^{th} place, and 3^{rd} place in 2017, 2018, and 2019, respectively. Our full framework CGS is in the top 3 over 4 teams achieving 0.75 in terms of global score in all three years. Our method also maintains the best stable and recall performance among the leading submissions.

In the future, we plan to consider modeling the semantic relationship among object instances in the segmentation process. We will also investigate Capsule-inspired [19, 51, 52, 79], and attention-inspired [5, 11, 12, 29] network architectures for better segmentation performance. We also aim to extend our work to camouflage analysis [24, 27, 75] in the near future.

Acknowledgment

This research is funded by Gia Lam Urban Development and Investment Company Limited, Vingroup, supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19, and National Science Foundation (NSF) under Grant No. 2025234. The first author would like to thank JSPS KAKENHI Grants (JP16H06302, JP18H04120, JP21H04907, JP20K23355, JP21K18023), JST CREST Grants (JPMJCR20D3, JPMJCR18A6). We also thank NVIDIA and AIOZ Pte Ltd for the support of GPU and computing infrastructure.

References

1. B. D. Brabandere, D. Neven, and L. V. Gool. Semantic instance segmentation for autonomous driving. In *CVPR Workshops*, 2017.
2. S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017.
3. H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018.
4. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.
5. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.
6. C. Chang and C. Lin. LIBSVM: A library for support vector machines. *Transactions on Intelligent Systems and Technology* , 2(3), 2011.
7. L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 2018.
8. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.
9. J. Cheng, S. Liu, Y.-H. Tsai, W.-C. Hung, S. Gupta, J. Gu, J. Kautz, S. Wang, and M.-H. Yang. Learning to segment instances in videos with spatial propagation network. *CVPR Workshops*, 2017.
10. F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

11. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.
12. B. Duke, A. Ahmed, C. Wolf, P. Aarabi, and G. W. Taylor. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *CVPR*, 2021.
13. M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2), 2010.
14. P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.
15. M. Fiaz, A. Mahmood, and S. K. Jung. Video object segmentation using guided feature and directional deep appearance learning. *CVPR Workshops*, 2020.
16. H. Guo, W. Wang, G. Guo, H. Li, J. Liu, Q. He, and X. Xiao. An empirical study of propagation-based methods for video object segmentation. *CVPR Workshops*, 2019.
17. P. Guo, L. Zhang, H. Zhang, X. Liu, H. Ren, and Y. Zhang. Adaptive video object segmentation with online data generation. *CVPR Workshops*, 2018.
18. K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017.
19. G. E. Hinton, A. Krizhevsky, and S. D. Wang. Transforming auto-encoders. In *International conference on artificial neural networks*, pages 44–51, 2011.
20. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.
21. J. Ji, S. Buch, A. Soto, and J. C. Niebles. End-to-end joint semantic segmentation of actors and actions in video. In *ECCV*, 2018.
22. A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. *CVPR Workshops*, 2017.
23. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.
24. T.-N. Le, Y. Cao, T.-C. Nguyen, M.-Q. Le, K.-D. Nguyen, T.-T. Do, M.-T. Tran, and T. V. Nguyen. Camouflaged instance segmentation in-the-wild: Dataset and benchmark suite. *ArXiv Pre-print: 2103.17123*, 2021.
25. T.-N. Le, K.-T. Nguyen, M.-H. Nguyen-Phan, T.-V. Ton, T.-A. Nguyen, X.-S. Trinh, Q.-H. Dinh, V.-T. Nguyen, A.-D. Duong, A. Sugimoto, T. V. Nguyen, and M.-T. Tran. Instance re-identification flow for video object segmentation. *CVPR Workshops*, 2017.
26. T.-N. Le, K.-T. Nguyen, M.-H. Nguyen-Phan, V. Ton-That, T.-A. Nguyen, X.-S. Trinh, Q.-H. Dinh, V.-T. Nguyen, A. D. Duong, A. Sugimoto, T. V. Nguyen, and M.-T. Tran. Instance re-identification flow for video object segmentation. *CVPR Workshops*, 2017.
27. T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto. Anabranch network for camouflaged object segmentation. *Journal of Computer Vision and Image Understanding*, 184:45–56, 2019.
28. T.-N. Le, T. V. Nguyen, Q.-C. Tran, L. Nguyen, T.-H. Hoang, M.-Q. Le, and M.-T. Tran. Interactive video object mask annotation. In *AAAI*, 2021.
29. T.-N. Le, A. Sugimoto, S. Ono, and H. Kawasaki. Attention r-cnn for accident detection. In *IEEE Intelligent Vehicles Symposium*, 2020.
30. Y. J. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *IJCV*, 114(1), 2015.
31. X. Li and C. C. Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. *CVPR Workshops*, 2018.
32. X. Li, Y. Qi, Z. Wang, K. Chen, Z. Liu, J. Shi, P. Luo, C. C. Loy, and X. Tang. Video object segmentation with re-identification. *CVPR Workshops*, 2017.
33. A. Lin, Y. Chou, and T. Martinez. Flow adaptive video object segmentation. *CVPR Workshops*, 2018.
34. 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.
35. D. Liu, D. Yu, M. Dong, L. Ma, J. Shao, J. Wang, C. Wang, and P. Zhou. An effective multi-level backbone for video object segmentation. *CVPR Workshops*, 2020.
36. N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016.
37. N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. In *CVPR*, 2015.
38. J. Luiten, P. Voigtlaender, and B. Leibe. Premvos: Proposal-generation, refinement and merging for the davis challenge on video object segmentation. *CVPR Workshops*, 2018.
39. J. Luiten, P. Voigtlaender, and B. Leibe. Combining premvos with box-level tracking for the 2019 davis challenge. *CVPR Workshops*, 2019.
40. G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017.

41. A. Newswanger and C. Xu. One-shot video object segmentation with iterative online fine-tuning. *CVPR Workshops*, 2017.
42. K. Nguyen, K. Nguyen, D. Le, D. A. Duong, and T. V. Nguyen. YADA: you always dream again for better object detection. *Multim. Tools Appl.*, 78(19):28189–28208, 2019.
43. K. Nguyen, K. Nguyen, D. Le, D. A. Duong, and T. V. Nguyen. You always look again: Learning to detect the unseen objects. *J. Vis. Commun. Image Represent.*, 60:206–216, 2019.
44. S. W. Oh, J. Lee, N. Xu, and S. J. Kim. A unified model for semi-supervised and interactive video object segmentation using space-time memory networks. *CVPR Workshops*, 2019.
45. F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.
46. 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:1704.00675*, 2017.
47. A. Robinson, F. J. Lawin, M. Danelljan, and M. Felsberg. Discriminative learning and target attention for the 2019 davis challenge on video object segmentation. *CVPR Workshops*, 2019.
48. A. Robinson, F. J. Lawin, M. Danelljan, F. S. Khan, and M. Felsberg. Learning fast and robust target models for video object segmentation. In *CVPR*, June 2020.
49. C. Rother, V. Kolmogorov, and A. Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *Transactions on Graphics*, 23(3), 2004.
50. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3), 2015.
51. S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. *NeurIPS*, 2017.
52. S. Sabour, A. Tagliasacchi, S. Yazdani, G. E. Hinton, and D. J. Fleet. Unsupervised part representation by flow capsules. *arXiv preprint arXiv:2011.13920*, 2020.
53. H. Seong, J. Hyun, and E. Kim. A kernel-based approach for video object segmentation. *CVPR Workshops*, 2020.
54. A. Shaban, A. Firl, A. Humayun, J. Yuan, X. Wang, P. Lei, N. Dhanda, B. Boots, J. M. Rehg, and F. Li. Multiple-instance video segmentation with sequence-specific object proposals. *CVPR Workshops*, 2017.
55. G. Sharir, E. Smolyansky, and I. Friedman. Video object segmentation using tracked object proposals. *CVPR Workshops*, 2017.
56. J. Sun, D. Yu, Y. Li, and C. Wang. Mask propagation network for video object segmentation. *CVPR Workshops*, 2018.
57. M.-T. Tran, T. Hoang, T. V. Nguyen, T.-N. Le, E. Nguyen, M. Le, H. Nguyen-Dinh, X. Hoang, and M. N. Do. Multi-referenced guided instance segmentation framework for semi-supervised video instance segmentation. *CVPR Workshops*, 2020.
58. M.-T. Tran, T.-N. Le, T. V. Nguyen, V. Ton-That, T.-H. Hoang, N.-M. Bui, T.-L. Do, Q.-A. Luong, V.-T. Nguyen, D. A. Duong, and M. N. Do. Guided instance segmentation framework for semi-supervised video instance segmentation. In *CVPR Workshops*, 2019.
59. M.-T. Tran, V. Ton-That, T.-N. Le, K.-T. Nguyen, T. V. Ninh, T.-K. Le, V.-T. Nguyen, T. V. Nguyen, and M. N. Do. Context-based instance segmentation in video sequences. *CVPR Workshops*, 2018.
60. Y. H. Tsai, M. H. Yang, and M. J. Black. Video segmentation via object flow. In *CVPR*, 2016.
61. P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019.
62. P. Voigtlaender and B. Leibe. Online adaptation of convolutional neural networks for the 2017 davis challenge on video object segmentation. *CVPR Workshops*, 2017.
63. V.Petrosyan, O. Örnsberg, and A. Proutiere. Video object segmentation via tracking edges and classifying segments. *CVPR Workshops*, 2018.
64. T. Vu-Le, H. Nguyen-Le, E. Nguyen, M. N. Do, and M. Tran. Video object segmentation with memory augmentation and multi-pass approach. *CVPR Workshops*, 2020.
65. B. Wang, C. Zheng, N. Wang, S. Wang, X. Zhang, S. Liu, S. Gao, K. Lu, D. Zhang, L. Shen, Y. Wang, and Y. Xu. Object-based spatial similarity for semi-supervised video object segmentation. *CVPR Workshops*, 2019.
66. Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019.
67. P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *ICCV*, 2013.
68. T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.
69. H. Xie, Y. Huang, A. Xu, J. Lan, and W. Sun. Depth-aware space-time memory network for video object segmentation. *CVPR Workshops*, 2020.

70. Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun. Upsnet: A unified panoptic segmentation network. In *CVPR*, 2019.
71. K. Xu, L. Wen, G. Li, L. Bo, and Q. Huang. Spatiotemporal cnn for video object segmentation. In *CVPR*, 2019.
72. N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang. Deep grabcut for object selection. *BMVC*, 2017.
73. S. Xu, L. Bao, and P. Zhou. Class-agnostic video object segmentation without semantic re-identification. *CVPR Workshops*, 2018.
74. S. Xu, D. Liu, L. Bao, W. Liu, and P. Zhou. Mhp-vos: Multiple hypotheses propagation for video object segmentation. In *CVPR*, 2019.
75. J. Yan, T.-N. Le, K.-D. Nguyen, M.-T. Tran, T.-T. Do, and T. V. Nguyen. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300, 2021.
76. J. Yang, B. Price, S. Cohen, H. Lee, and M. H. Yang. Object contour detection with a fully convolutional encoder-decoder network. In *CVPR*, 2016.
77. Z. Yang, Y. Ding, Y. Wei, and Y. Yang. Cfbi+: Collaborative video object segmentation by multi-scale foreground-background integration. *CVPR Workshops*, 2020.
78. P. Zhang, L. Hu, B. Zhang, and P. Pan. Spatial constrained memory network for semi-supervised video object segmentation. *CVPR Workshops*, 2020.
79. W. Zhang, P. Tang, and L. Zhao. Remote sensing image scene classification using cnn-capsnet. *Remote Sensing*, 11(5):494, 2019.
80. H. Zhao. Some promising ideas about multi-instance video segmentation. *CVPR Workshops*, 2017.
81. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017.
82. B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 2017.