OVSNet : Towards One-Pass Real-Time Video Object Segmentation

Peng Sun[†*] Peiwen Lin[‡*] Guangliang Cheng[‡] Jianping Shi[‡] Jiawan Zhang[♭] Xi Li[†]

[†]Zhejiang University [‡]SenseTime Research [♭]Tianjin University

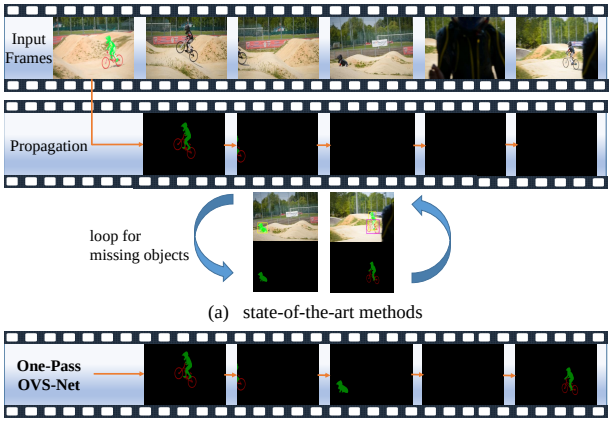
{sunpeng1996, xilizju}@zju.edu.cn jwzhang@tju.edu.cn
{linpeiwen, chengguangliang, shijianping}@sensetime.com

Abstract

*Video object segmentation aims at accurately segmenting the target object regions across consecutive frames. It is technically challenging for coping with complicated factors (e.g., shape deformations, occlusion and out of the lens). Recent approaches have largely solved them by using back-forth re-identification and bi-directional mask propagation. However, their methods are extremely slow and only support offline inference, which in principle cannot be applied in real time. Motivated by this observation, we propose a new detection-based paradigm for video object segmentation. We propose an unified One-Pass Video Segmentation framework (OVS-Net) for modeling spatial-temporal representation in an end-to-end pipeline, which seamlessly integrates object detection, object segmentation, and object re-identification. The proposed framework lends itself to one-pass inference that effectively and efficiently performs video object segmentation. Moreover, we propose a mask-guided attention module for modeling the multi-scale object boundary and multi-level feature fusion. Experiments on the challenging DAVIS 2017 demonstrate the effectiveness of the proposed framework with comparable performance to the state-of-the-art, and the great efficiency about **11.5 fps** towards pioneering real-time work to our knowledge, more than **5 times** faster than other state-of-the-art methods.*

1. Introduction

As an important and challenging problem in computer vision, video object segmentation is typically cast as a problem of instance-aware pixel-wise classification across consecutive frames with the object mask provided in the first frame. In principle, video object segmentation requires accomplishing the following two tasks: 1) intra-frame object instance segmentation; and 2) inter-frame object instance association. With the single first frame label, how to effectively model spatio-temporal context is key issue for robust video object segmentation.



(a) state-of-the-art methods

(b) Our OVS-Net: Make online possible.

Figure 1. We focus on the person and the bicycle in the original input frames. As shown in (a), the state-of-the-art methods such as VS-ReID or DyeNet need traverse the video many times to retrieve the missing objects due to occlusion or out of the lens. Our OVS-Net can produce accurate segmentation results in a one-pass manner, as shown in (b). Best viewed in color.

Traditional methods [15, 27, 36, 5, 10] are graph based to build spatio-temporal graphs for contextual relationship in video object segmentation. With the power of deep learning, such spatial-temporal correlation is usually carried out within a deep neural network framework, which takes an end-to-end convolution neural network learning architecture over the benchmark datasets (e.g., DAVIS Challenge [31, 30]). One stream of spatial-temporal modeling is via optical flow across frames [33, 13, 12]. Such optical

[*]equal contribution

flow based methods [13, 33, 12] are prone to errors at occluded or reappearing objects, which commonly appears in real world. In order to address the problem of object occlusion, vanishing, or reappearing, another stream of methods [18, 20, 14] utilize object re-identification (Re-ID) techniques [35] where the occlusion or reappearing problem is largely solved by the powerful Re-ID features. However, as shown in Figure 1(a), the aforementioned approaches [18, 20] essentially adopt an offline Re-ID strategy (back-forth Re-ID) that traverses the whole videos for several times and then performs bi-directional mask propagation during inference, which is far away from the real-time or online requirements.

To enjoy the ability to handle occlusion or reappearing problems and towards real-time application, we propose to build a detection-based pipeline as well as an effective unified end-to-end framework named OVS-Net for modeling spatial-temporal representation. The proposed framework takes a seamless integration of object detection, object segmentation, and object re-identification. Specifically, the proposed framework carries out the joint learning procedure of object segmentation and object Re-ID during training.

To enable efficient inference and associate the corresponding outputs by OVS-Net across frames, we propose an one-pass strategy to produce accurate segmentation results frame by frame in a one-pass manner with online Re-ID, as shown in Figure 1(b). The one-pass strategy is based on a cascaded rule consisting of three paths: IOU path, Re-ID path, and flow path, resulting in great computational efficiency. Furthermore, we propose a mask-guided attention module to fully explore the multi-scale object information and multi-level feature fusion, which leads to performance improvements in object boundary. As a matter of fact, the joint training aims at enabling the network to learn the capability of object segmentation within individual frames and object Re-ID association across different frames. In contrast, the motivation of the cascaded inference strategy is to ensure object segmentation towards the real-time performance.

Specifically, the instance masks generated in [18, 20] are warped from previous frames using optical flow and then refined to get better predictions. Instead, our method generates masks independently for each frame and then links these masks through the proposed one-pass cascaded inference strategy. The pipeline we propose can be seen as a new detection-based paradigm different against their propagation-based methods [18, 20]. Our approach achieves a comparable performance to the state-of-the-art approaches and can perform *11.5 frames per seconds*. Our OVS-Net is about *5 times* faster than DyeNet [18], *35 times* faster than VS-ReID [20]. Extensive experiments demonstrate the performance effectiveness and efficiency of this work above against the state-of-the-art.

The main contributions of this work can be summarized as follows.

- We propose the first end-to-end detection-based pipeline for video object segmentation as well as a joint training framework for modeling spatio-temporal representation, which seamlessly integrates the modules of object detection, segmentation, and re-identification in an end-to-end manner.
- We propose an effective one-pass cascaded inference strategy, which leads the network inference towards real-time performance (around *11.5 fps*) in a one-pass manner with online Re-ID.
- Furthermore, we propose a mask-guided attention module to fully explore the multi-scale object information within a feature pyramid network, which leads to performance improvements in object boundary.

2. Related Work

Semantic Segmentation Video object segmentation pays great attention to the quality of segmentation results. With the boom of deep learning, semantic segmentation [2, 25, 37, 3, 38] have shown extremely good performance in static images. Different from semantic segmentation which has pre-defined semantic label, in video object segmentation, we focus on the class-agnostic object segmenation.

Instance Segmentation In video object segmentation, we also need to distinguish between different instances. There are many effective methods in instance segmentation [6, 7, 8, 21, 24], and Mask R-CNN is the most popular solution recently. Our framework is built on Mask R-CNN and improves it from many aspects, which is the first effective end-to-end detection-based framework for video object segmentation.

Video Object Segmentation Recently, deep learning based model is the most promising one to tackle video object segmentation task. For example, Perazzi et al. [29] introduced video object segmentation as a guided instance segmentation problem via frame-by-frame pixel-labeling strategy. Caelles et al. [1] introduced one transfer learning based algorithm (OSVOS) to tackle the task of foreground segmentation. While OSVOS can achieve impressive performance, it is slightly sensitive to large changes in object appearance. To tackle this limitation, Voigtlaender et al. [34] proposed an online adaptive video object segmentation algorithm that can update the network online using the selected training examples. Tsai et al. [33] considered the video segmentation through a optical flow based algorithm to maintain object boundaries and temporal consistency simultaneously. Li et al. [17] incorporated the neighborhood

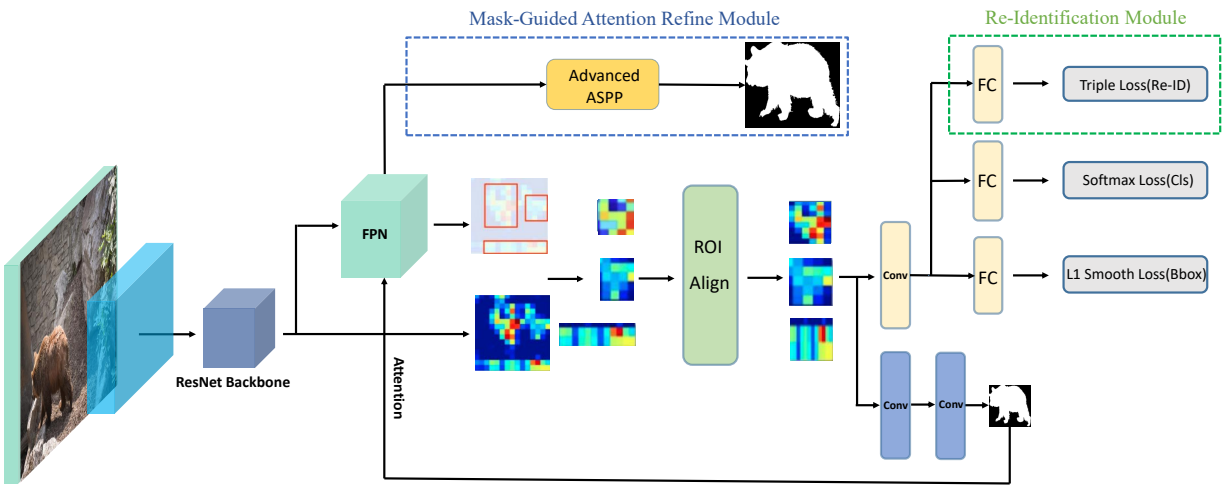


Figure 2. Illustration of our framework. The Re-ID branch is our online Re-Identification module and the middle part of the above is the Mask-guided Attention module which use the Mask R-CNN prediction as the attention map and combine the muti-level features of FPN for producing more accuracy segmentation result than original Mask R-CNN.

reversible flow to segment the foreground objects and suppress the distractors. Video propagation networks [12] incorporated temporal bilateral network and adaptive filtering strategy to propagate information forward to the future frames. However, they are inability to track the object that re-appears in the video.

To tackle the above limitation, some state-of-the-art approaches [14, 18, 20] incorporated person re-identification algorithm [35] into the video object segmentation framework. For instance, Instance Re-Identification Flow (IRIF) [14] can track and detect the re-appeared instance via the instance re-identification module and mask propagation module. Li et al. [20] adapted Re-ID approach and a two-stream mask propagation model in their framework, while it is slow in inference and slightly sensitive to the pose variations. To overcome these shortcomings, Li et al. [18] proposed a substantially robust and efficient network with the attention mechanism, while their back-forth Re-ID and bi-directional propagation methods are offline and takes about 0.43 seconds per frame, which is far from real-time application. Meanwhile, different from their propagation-based methods [18, 20], we propose the first end-to-end detection-based pipeline, in this way, we do not need to use optical flow information during the training process, which simplifies the training process. In addition, the attention mechanism proposed in [18] is in propagation process, while our attention module is for better modeling the intra-frame spatial context information. Furthermore, although [18] conducted one-iteration Re-ID experiment, they still adopted an offline back-forth greedy strategy (offline sorted for Re-ID and bi-directional propagation), which cannot be used to deal with online or streaming videos. Though our proposed detection-based pipeline, we can perform online Re-ID and achieve comparable performance, meanwhile our OVS-Net is about *5 times* faster than DyeNet [18].

3. Framework

In this section, we first introduce our effective detection-based unified framework during training phase, for intra-frame spatial context modeling and inter-frame temporal object association in an end-to-end scheme by describing two sub-modules: Mask-Guided Attention (MGA) module and Re-ID module. Later we describe our one-pass cascaded strategy to perform object segmentation within individual frames and sequentially link the corresponding ID-specific object segmentation masks frame by frame in a one-pass manner towards real-time.

3.1. Fully End-to-End Network

Given a video and the target object masks in the first frame, our goal is to segment the target instance objects in future frames. The proposed end-to-end framework seamlessly integrates the object detection, object segmentation and object re-identification. Specifically, it is the first detection-based pipeline for video object segmentation. As illustrated in Figure 2, our framework consists of Mask R-CNN [8], online Re-Identification (Re-ID) module and MGA module. Specifically, the MGA module utilizes an attention strategy to fuse multi-level information that can greatly improve the object boundary performance, for better intra-frame spatial context modeling. The Re-ID module

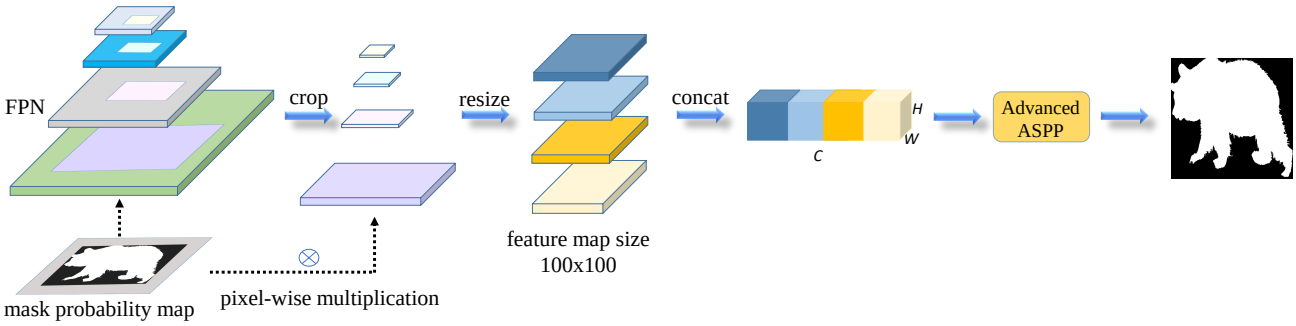


Figure 3. Illustration of the Mask-Guided Attention Module.

is capable of retrieving the reappeared objects and is incorporated into the Mask R-CNN [8] network, which can adapt to each other during the joint optimization, for inter-frame temporal object association.

3.1.1 Mask-Guided Attention Module

First, we model the intra-frame spatial context information for each image. We use Mask R-CNN [8] as our base instance segmentation modeling method. Generally, the output size of Mask R-CNN is 28×28, which is relatively small that can not achieve satisfactory aligned boundary well, especially for big objects. To fully explore the multi-scale object information, a powerful module named Mask-Guided Attention (MGA) is proposed, which use the prediction of Mask R-CNN as the attention map and combine the muti-level features of FPN [22] to produce bigger and more accurate segmentation result than the original output of Mask R-CNN without introducing a lot of time overhead.

Different from original Mask R-CNN head, here we use the MGA module to predict a class-agnostic foreground/background mask. In addition, the MGA module will expand the attention masks and features at the corresponding position by 20% (for the attention mask, we fill it with a fixed probability value of 0.5) in order to segment the boundary of the object accurately. We use the probability map of the output of Mask R-CNN as the attention map, crop the feature maps of FPN according to the coordinate of enlarged bounding bboxes, and pixel-wise multiply the cropped feature maps with the attention map, then resize the feature maps to 100x100 and pass the last block of ResNet-101 [9] followed by the advanced aspp module to get the final segmentation result. The loss function is the pixel-wise cross entropy loss and the gradient generated by MGA module will only propagate backward to itself, and will not affect FPN and Mask R-CNN. Figure 3 illustrates the Mask-guided Attention module. For the Advanced ASPP module, we simply combine the ASPP module of DeepLabv3 [3] and the PPM module of PSPNet [37], which bring a slight improvement in performance.

3.1.2 Online Re-identification Module

Since we have modeled the intra-frame spatial context information, now we focus on inter-frame object instance association. We propose the powerful online re-identification module to find the reappeared objects in real time. In all experiments, we assign a unique, non-repeating re-identification label for each target object in each video. We use the Triplet-Loss [32] for training the online re-identification module, we project the features extracted by RoI-Align Pooling into a 128-dimensional feature vector by simply adding two fully connected (fc) layers and the last *fc* layer is followed by a dropout layer with a dropout rate 0.2. In principle, our training framework adopts a multi-task learning pipeline that jointly optimizes the Re-ID module in conjunction with the Mask-RCNN module on the basis of the same backbone network. The joint training process is carried out for each minibatch consisting of different frames from the same or different videos. For the Re-ID training, we select N bbox proposals in each frame, and then get its Re-ID feature through the Re-ID module. Suppose each minibatch has M frames, as a result, we totally have M*N training examples from diverse (the same or different) frames, enough to use the triplet loss. The Re-ID training aims to learn object association for the between-frame training examples, while the Mask-RCNN training seeks for learning the spatial segmentation for the within-frame training examples. Finally, the learning gradients of the above two modules are simultaneously back-propagated to the shared backbone network for the joint optimization process.

The loss function of the online Re-ID module is as follows,

$$L_{reid}=\sum_{i=1}^{N}\left[\|f(x_i^a)-f(x_i^p)\|_2^2-\|f(x_i^a)-f(x_i^n)\|_2^2+\alpha\right]_+, \quad (1)$$

where p denotes positive object set that has the same Re-ID label with object a, and n denotes positive object set that has different Re-ID labels with object a. $f(x_i^a)$ is the reid-feature for object a, and $f(x_i^p)$ is the reid-feature for the positive object that is the most dissimilar to the object a, and $f(x_i^n)$ is the reid-feature for the negative object that is most similar to the object a. α is the margin of triplet-loss and we experimentally set it to 1.0.

3.2. One-Pass Cascaded Inference Strategy

With our well-designed inference strategy, the video frames need to traverse only once rather than many times, which makes it possible for the video object segmentation in real time. In principle, the cascaded inference strategy makes full use of smooth transition across adjacent frames to locate the object segmentation results by lightweight optical flow, while long-term inter-frame objects after occlusion or drastic motion can be handled by Re-ID. The object association rules are based on a cascaded strategy consisting of three paths: IOU path, Re-ID path and flow path. We define ρ_{reid}, ρ_{iou}, C_{reid}^n as the similarity threshold for reid-feature, the threshold for the IOU and the reid-feature collection for *n-th* instance, respectively. So the number of collection C_{reid} is equal to the number of instances need to track in the video. As shown in Figure 4, given the mask of last frame M_{i-1} and the candidate objects P_i predicted by our OVS-Net, we aim to generate the probability map for certain instance in current frame M_i by three cascaded inference paths, which is IOU-Path, Reid-Path, and Flow-Path.

To better explain above three paths, we choose the video with only one instance as our example and denote this instance as I. If there are multiple instances in the video, the same operation will be simply executed for all instances. Before executing these three operation, we first warp the mask of last frame for instance I by optical flow extracted by FlowNet 2.0 [11] as the mask template M_{i-1}^{warp} for current frame and then generate the candidate segmentation result P_i and its reid-feature in current frame by our OVS-Net. After that, the following three operation will be executed:

IOU-Path: As mentioned above, we have obtained the candidate segmentation results for current frame and its reid-features. In this operation, we first calculate the IOU between all candidates and the instance I in the mask template M_{i-1}^{warp}. If there exist candidates whose IOU are larger than ρ_{iou}, we take the candidate that has the largest IOU as the final mask for the instance I in current frame and also save its reid-feature in the instance I's reid-feature collection C_{reid}^I, otherwise, the program will execute the next operation Reid-Path.

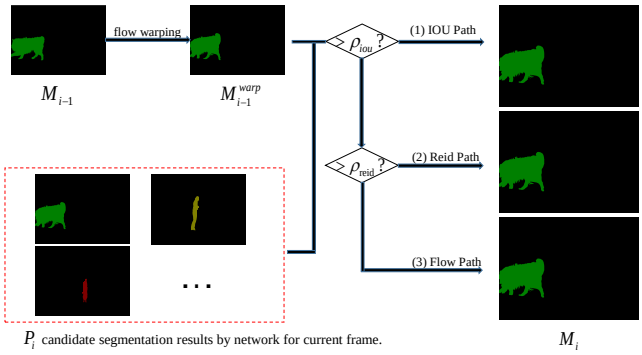


Figure 4. The process of our one-pass cascaded inference strategy. **Best viewed in color.**

Reid-Path: When the IOU-Path setting can't be met, it is possible that we have lost instance I in previous frame, so we need to use I's reid-feature saved in C_{reid}^I to retrieve instance I in current frame. Firstly, The similarity between all candidate reid-features and the instance I's reid-features saved in C_{reid}^I will be calculated through measuring the euclidean distance. If there exist candidates whose similarity with at least 30% reid-features in collection C_{reid}^I is smaller than ρ_{reid}, we take the candidate who has the smallest similarity with instance I as its final mask in current frame.

Flow-Path: If no any candidates satisfy the paths above, a rough mask warped by optical flow M_{i-1}^{warp} will be sent to MGA module to generate the final mask for instance I in current frame.

4. Experiments

4.1. Implementation Details

We re-implement Mask R-CNN and FPN based on Pytorch 0.3.1 [28] and use ResNet-50 [9] as our backbone. We train the entire network starting from pre-trained ImageNet [4] weights on COCO [23] dataset. Follow the common configurations, we use the muti-scale training, the shorter edges of the input images are 640, 720, 800, 920, 1080, the max edge of the images is 1900 and the anchor scales are 3, 5, 8. After pretrained on COCO, the origin Mask R-CNN is first trained on DAVIS training sets for two epochs in order to initialize the whole Network. Then the whole network with Re-ID module and the Mask-guided Attention module is trained on DAVIS for three epochs. We fix a mini-batch size of 32 images for Mask R-CNN, momentum 0.9 and weight decay of 10^{-4}, and for every image in a mini-batch, we select two proposals which have the largest IoU with ground-truth bboxes for the MGA Module and Re-ID module. So the mini-batch of the Re-ID mod-

ule and the MGA Module are both 64. The initial learning is 0.04 and dropped by a factor of 10 after each epoch. For data augmentation, we employ the Lucid Data Dreaming [13] to generate augmented images using the first frame on testing videos and add them into the training set that adopts our model to the target video domain. The overall loss function of our model is formulated as:

$$L_{total} = L_{mask} + L_{loc} + L_{cls} + L_{reid} + \lambda L_{refine} \quad (2)$$

The L_{reid} is the triplet loss of Re-ID module. L_{loc}, L_{cls}, L_{mask} respectively represent location loss, classification loss and segmentation loss of the Mask R-CNN. The L_{refine} is the pixel-wise cross-entropy loss of the MGA module where λ is a weight that balances these loss terms for better muti-task learning. During inference phase, we only use the single scale without muti-scale and horizontal flip testing and we keep the proposals with a score greater than 0.05 and perform non-maximum suppression which have an IOU of 0.6 with a proposal with a higher score.

4.2. Benchmark

Datasets In order to verify the effectiveness and robustness of our method, we evaluate our method on DAVIS$_{16}$ [30], DAVIS$_{17}$ [31] and SegTrack$_{v2}$ [16] datasets. The DAVIS dataset is a public dataset, benchmark, and competition specially designed for the task of video object segmentation, spanning multiple occurrences of common video object segmentation challenges such as occlusions, fast-motion, appearance changes and out of the lens. SegTrack$_{v2}$ dataset is a video segmentation dataset with full pixel-level annotations on multiple objects in each frame within each video. We conduct a complete ablation study on the DAVIS$_{17}$ test-dev dataset. In this section, we compare the first detection-based one-pass OVS-Net with other existing state-of-the-art methods and show it can achieve the comparable performance towards real-time on these standard datasets.

Evaluation Metric For DAVIS$_{17}$ Dataset, we follow [31] that employs the Jaccard index *J* defined as the *intersection-over-union* of the estimated segmentation and the groundtruth mask and employ Contour Accuracy *F* to compute the contour-based precision and recall and their average *G* measures for evaluation. For DAVIS$_{16}$ and SegTrack$_{v2}$ Datasets, we use the Jaccard index *J* defined as the *intersection-over-union* across all instances to evaluate the performance, same as other methods.

Different testing patterns For testing phase, there are two patterns, since the first frame annotations are provided as described in [18]. They can be further divided into **per-dataset** and **per-video** finetuning. In per-dataset finetuning, we merge all first frame annotations from test set into training set that adopts the model to the target video domain to obtain a dataset-specific model. However, per-video finetuing means that finetune a model on each testing video, i.e., the number of final models is as many as the number of videos during the test phase. Obviously, the former has better universality. Table 1 lists the *J*-means, *F*-means and *G*-means on DAVIS$_{17}$ *test-dev* dataset with other existing state-of-the-art methods. And for DAVIS$_{16}$, we use its own train set and its first frames in val set to train our model and evaluate the model on DAVIS$_{16}$ val set. For SegTrack$_{v2}$, we only use the first frames of the videos as the training data and finetune the model which is pretrained on DAVIS$_{17}$. And Table 2 lists the performance on DAVIS$_{16}$ val set and SegTrack$_{v2}$ dataset. Due to memory limitation, the backbone used in our all experiments is ResNet50 [9].

Method	*J*-mean	*F*-mean	*G*-mean	per-video
OSVOS [1]	47.0	54.8	50.9	✓
OSVOS-S [26]	52.9	62.1	57.5	✓
OSAVOS [34]	53.4	59.6	56.5	✓
LucidTracker* [13]	60.1	68.3	64.2	✓
DyeNet [18]	65.8	70.5	68.2	×
VS-ReID* [20]	63.3	67.0	65.2	×
OVS-Net*	62.5	68.4	65.5	×

Table 1. Results on DAVIS$_{17}$ test-dev dataset. * means only using the single scale for testing and per-video means whether to use the per-video finetuning.

Method	Davis$_{16}$	SegTrack$_{v2}$	per-video
OSVOS [1]	79.8	65.4	✓
MSK [29]	80.3	70.3	✓
OSAVOS [34]	85.7	-	✓
LucidTracker [13]	84.8	77.6	✓
DyeNet [18]	84.7	78.7	×
OVS-Net	84.6	76.9	×

Table 2. Results on DAVIS$_{16}$ dataset and SegTrackv2 datasets. Per-video means whether to use the per-video finetuning.

4.3. Ablation Study

In this section, in order to verify the effectiveness of each component of our framework, we perform the ablation study experiments. All performance are reported on the *test-dev* set of DAVIS$_{17}$.

Robustness of the Online Re-Identification Module

For Re-ID module, we performed a series of experiments with different ρ_{reid} value in Re-ID path or without the Re-ID module. The ρ_{reid} is extremely important for recall and precision. Table 3 shows the different ρ_{reid} values with their performance. And using the ρ_{reid} = 2.3, our OVS-

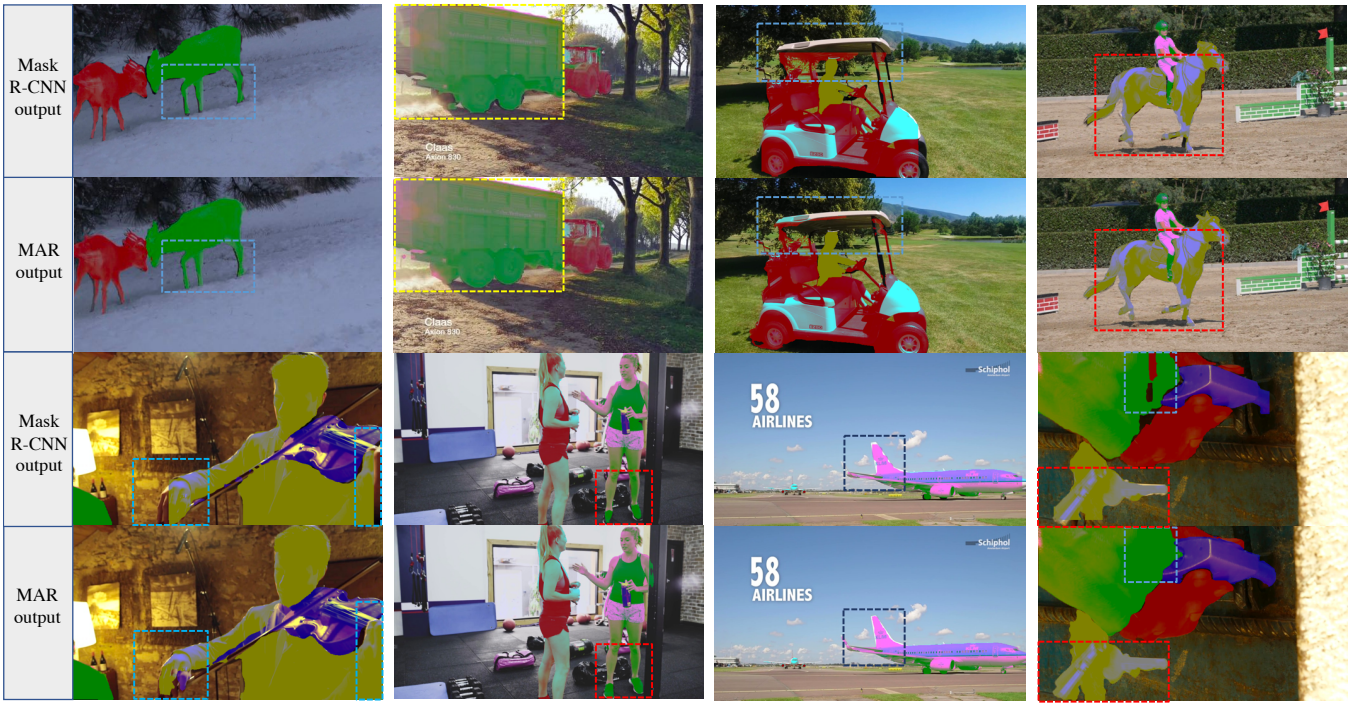


Figure 5. Visualization results with or without the Mask-guided Attention Module. The following line is the result with the Mask-guided Attention Module, having better segmentation result. The dashed box is drawn for easy comparison. **Best viewed in color.**

Net achieves the best performance. In the following experiments, we fix the ρ_{reid} as 2.3.

Methods	*J*-mean	*F*-mean	*G*-mean
OVS-Net(without Re-ID)	53.1	56.7	54.9
OVS-Net(with ρ_{reid} = 2.1)	56.9	63.7	60.3
OVS-Net(with ρ_{reid} = 2.2)	57.6	63.2	60.4
OVS-Net(with ρ_{reid} = 2.3)	**58.3**	**63.7**	**61.0**
OVS-Net(with ρ_{reid} = 2.4)	57.4	63.2	60.3
OVS-Net(with ρ_{reid} = 2.5)	57.0	63.0	60.0

Table 3. Ablation study for Re-ID module with different ρ_{reid} value on DAVIS_{17} test-dev dataset.

Effectiveness of the Mask-guided Attention Module

The Mask-guided Attention Module is very effective, to prove this, we performed a series of ablation study experiments. Our network is essentially multi-task learning, and the loss weight between different tasks is very important in multi-task learning. Figure 5 shows that the segmentation result of our Mask-guided Attention Module is better than the original Mask R-CNN. And Table 4 lists the results with different loss weight λ. We can conclude that MGA module brings above 3 points performance than baseline OVS-Net. We find that the OVS-Net achieves the best performance when λ is 1.3.

Methods	*J*-mean	*F*-mean	*G*-mean
OVS-Net(without MGA)	58.3	63.7	61.0
OVS-Net(with λ = 1.0)	62.1	66.7	64.4
OVS-Net(with λ = 1.1)	63.2	67.4	65.3
OVS-Net(with λ = 1.2)	62.1	67.6	64.9
OVS-Net(with λ = 1.3)	**62.5**	**68.4**	**65.5**
OVS-Net(with λ = 1.4)	62.0	67.8	64.9

Table 4. Ablation study for Mask-guided Attention Module with DAVIS_{17} test-dev dataset.

Effectiveness of the Inference Strategy

We conduct a series of comparative experiments to carefully choose ρ_{iou}. The ρ_{iou} is also important in IOU path as mentioned before. In our cascaded strategy, the object association is almost solved in IOU path. Table 5 demonstrates the final performance at different values. We find that the OVS-Net achieves the best performance when the ρ_{iou} is 0.3.

Effectiveness of each module of our OVS-Net

Finally, we conduct a ablation study to evaluate each module of OVS-Net with DAVIS_{17} test-dev dataset. We add effective modules step by step and show their performance as shown in Table 6. All models in the experiments are end-to-end trained and only use the single scale (for

Figure 6. Visualization results of OVS-Net's prediction. The first column shows the first frame ground-truth masks of each video sequence. The frames are chosen randomly. **Best viewed in color.**

Methods	*J*-mean	*F*-mean	*G*-mean
OVS-Net(with $\rho_{iou} = 0.1$)	62.9	64.8	63.9
OVS-Net(with $\rho_{iou} = 0.2$)	61.3	67.4	64.4
OVS-Net(with $\rho_{iou} = 0.3$)	**62.5**	**68.4**	**65.5**
OVS-Net(with $\rho_{iou} = 0.4$)	61.7	62.9	62.3

Table 5. Ablation study for Inference Strategy with DAVIS_{17} test-dev dataset. ρ_{iou} means the threshold for the IOU mentioned before.

DAVIS_{17} test-dev, we fix the shorter edges of input images are 800) without muti-scale and horizontal flip testing in testing phase. OVS-Net with Re-ID module outperforms baseline OVS-Net by 6.1 points. To further introduce the MGA module into our framework, it can also brings another 4.5 points performance.

Methods	*G*-mean	Δ *G*-mean
OVS-Net(base)	54.9	-
+ Re-ID Module	61.0	+6.1
+MGA Module	65.5	+4.5

Table 6. Ablation study for each module of OVS-Net with DAVIS_{17} test-dev dataset.

4.4. Speed Analysis

In this section, we compare the speed of our model with the state-of-the-art methods. Our model is the first end-to-end detection-based pipeline without optical flow during the training phase, significantly reducing training time and training complexity. As far as we know, the full On-AVOS [34] takes roughly 13 seconds per frame and achieves 85.7 mIOU on DAVIS_{16} val dataset. The VS-ReID [20]'s speed is about 0.33FPS and the DyeNet [18] is quicker and their speed is about 2.4FPS without per-dataset finetuning, after per-dataset finetuning, their runing time is 0.43FPS. According to the above running time, these approaches are far from the real-time applications. Identical to the state-of-the-art DyeNet work [19], we use a single TitanXP GPU hardware and pytorch to perform evaluations for speed analysis throughout all the experiments. Moreover, we directly quote the results of time consumption and accuracy of other algorithms mentioned in [19] for fair comparison. As shown in Table 7, our OVS-Net achieves the 84.6 mIOU on DAVIS_{16} and the running time is **11.5FPS**.

5. Conclusion

We propose the first end-to-end detection-based pipeline for challenging video object segmentation, as well as a spatio-temporal training framework, which seamlessly integrates object detection, object segmentation, and object

Methods	G-mean	speed/per frame	FPS
OnAVOS [34]	85.7	13s	0.08FPS
VS-ReID [20]	-	3s	0.33FPS
DyeNet [18]	84.7	0.42s	2.4FPS
DyeNet [18]*	86.2	2.3s	0.43FPS
OVS-Net	84.6	**0.087s**	**11.5FPS**

Table 7. The inference speed analysis with DAVIS_{16} val dataset. * means after per-dataset finetuning using the first frame of testing dataset in their method.

re-identification. In principle, the proposed scheme is simple yet effective with a one-pass sequential inference strategy with online Re-ID, which ensures the computational efficiency of video object segmentation towards real-time performance (around **11.5 fps**). The mask-guided attention module is proposed to fully model the multi-scale object boundary information, which leads to performance improvements in object boundary. Extensive experiments demonstrate the performance effectiveness and efficiency of this work against the state-of-the-art.

References

[1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6

[2] L. 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. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 2

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 4

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009. 5

[5] M. Grundmann, V. Kwatra, M. Han, and I. A. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, pages 2141–2148, 2010. 1

[6] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, pages 297–312, 2014. 2

[7] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, pages 447–456, 2015. 2

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. 2, 3, 4

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 4, 5, 6

[10] S. Hickson, S. Birchfield, I. A. Essa, and H. I. Christensen. Efficient hierarchical graph-based segmentation of RGBD videos. In *CVPR*, pages 344–351, 2014. 1

[11] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 1647–1655, 2017. 5

[12] V. Jampani, R. Gadde, and P. V. Gehler. Video propagation networks. In *CVPR*, pages 3154–3164, 2017. 1, 2, 3

[13] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for multiple object tracking. *arXiv preprint arXiv:1703.09554*, 2017. 1, 2, 6

[14] 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. Tran2. Instance re-identification flow for video object segmentation. 2017. 2, 3

[15] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, pages 1995–2002, 2011. 1

[16] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013. 6

[17] J. Li, A. Zheng, X. Chen, and B. Zhou. Primary video object segmentation via complementary cnns and neighborhood reversible flow. In *ICCV*, pages 1426–1434, 2017. 2

[18] X. Li and C. C. Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. *arXiv preprint arXiv:1803.04242*, 2018. 2, 3, 6, 8, 9

[19] X. Li and C. C. Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *ECCV*, pages 93–110, 2018. 8

[20] X. Li, Y. Qi, Z. Wang, K. Chen, Z. Liu, J. Shi, P. Luo, X. Tang, and C. C. Loy. Video object segmentation with re-identification. *arXiv preprint arXiv:1708.00197*, 2017. 2, 3, 6, 8, 9

[21] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, pages 4438–4446, 2017. 2

[22] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017. 4

[23] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[24] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. *CoRR*, abs/1803.01534, 2018. 2

[25] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Deep learning markov random field for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(8):1814–1828, 2018. 2

[26] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. Video object segmentation without temporal information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 6

[27] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, pages 1777–1784, 2013. 1

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 5

[29] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, pages 3491–3500, 2017. 2, 6

[30] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 1, 6

[31] 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. 1, 6

[32] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4

[33] Y. Tsai, M. Yang, and M. J. Black. Video segmentation via object flow. In *CVPR*, pages 3899–3908, 2016. 1, 2

[34] P. Voigtlaender and B. Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *BMVC*, 2017. 2, 6, 8, 9

[35] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, pages 3376–3385, 2017. 2, 3

[36] C. Xu, C. Xiong, and J. J. Corso. Streaming hierarchical video segmentation. In *ECCV*, pages 626–639, 2012. 1

[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017. 2, 4

[38] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529–1537, 2015. 2