Adaptive Video Object Segmentation with Online Data Generation

Ping Guo[1,*] Lidan Zhang[1,*] Hangjian Zhang[2] Xiang Liu[3] Haibing Ren[1] Yimin Zhang[1]
[1]Intel Labs China [2]Technische Universität Braunschweig [3]University of Science and Technology Beijing
{ping.guo, lidan.zhang, haibing.ren, yimin.zhang}@intel.com
hangjian.zhang@tu-braunschweig.de
xiangliu0330@gmail.com

Abstract

The performance of deep learning models heavily depends on large amounts of labeled training data, while only one annotation is available in the one shot object segmentation task. To address this insufficient data problem, we propose an adaptive video object segmentation method with online data generation. Our adaptation process iteratively conducts two modules: instance segmentation and online data generation. In instance segmentation, the segmentation model is adapted to each frame by fine-tuning on online data from nearby frames. After all frames are processed in one iteration, the segmentation model is updated again to enhance its robustness by fine-tuning on all online data. In data generation, we first extract possible mask proposals in each frame. Then a joint tracking and re-identification method is proposed to filter and rank proposals in terms of temporal and appearance similarities. Finally, the highest confident proposal is chosen by merging with segmentation results and accumulated to update the above segmentation model. The effectiveness of the proposed method is demonstrated on DAVIS 2018 challenge dataset, with a region Jaccard of 67.5% and a boundary F measure of 71.5%.

1. Introduction

With the development of deep learning techniques, convolutional neural networks (CNN) have achieved state-of-the-art results in many computer vision tasks, including object detection, tracking, segmentation, etc. However, their dependence on a large amount of labeled training data prevents these deep models in one shot object segmentation where only one instance annotation is available. One promising work, the One-Shot Video Object Segmentation, OSVOS [1] fine-tuned a pre-trained network on the given one annotation. This schema is prone to overfitting with poor performance for either foreground or background variations.

To capture online variations, the method Online Adaptive Video Object Segmentation (OnAVOS) [2] was proposed by updating the OSVOS network online using selected outputs from the same network as training data. This "self-loop" schema introduces more training samples and slightly increased robustness on foreground appearance variation. The problem is the quality of the generated samples cannot be guaranteed, which brings error propagation and segmentation artifacts throughout the video. Another work used in the one shot segmentation task is called lucid data dreaming [5], which synthesized training data by simulating foreground change in illumination, deform, motion, etc. and in-painting on a dynamic background. However, their transformation is limited and still hard to cover the large variation in either foreground or background in unseen frames.

The proposed method in this paper aims to improve the performance by generating more reliable training samples with novel training schema for segmentation network. In data generation, for each frame, mask proposals were initialized with the outputs of mask branch in Mask R-CNN [4]. Further, the confident proposals were found and ranked by utilizing temporal and appearance consistency with a novel tracking and re-identification module. By merging with current segmentation results, the highest confident mask is chosen and append to our online training sample set. In training of the segmentation network, we use a two-level training schema to capture both appearance representation in one frame (local adaptation), and appearance variance in the whole video (global adaptation). Finally, we explore more synthesizing data by introducing external background images.

2. Proposed Method

As shown in Figure 1, our system for online object segmentation consists of two stages: offline-training and online adaptation. In offline training, the individual network of each task is independently trained with different datasets. The detection and segmentation nets are further

[*] the first two authors contributed equally

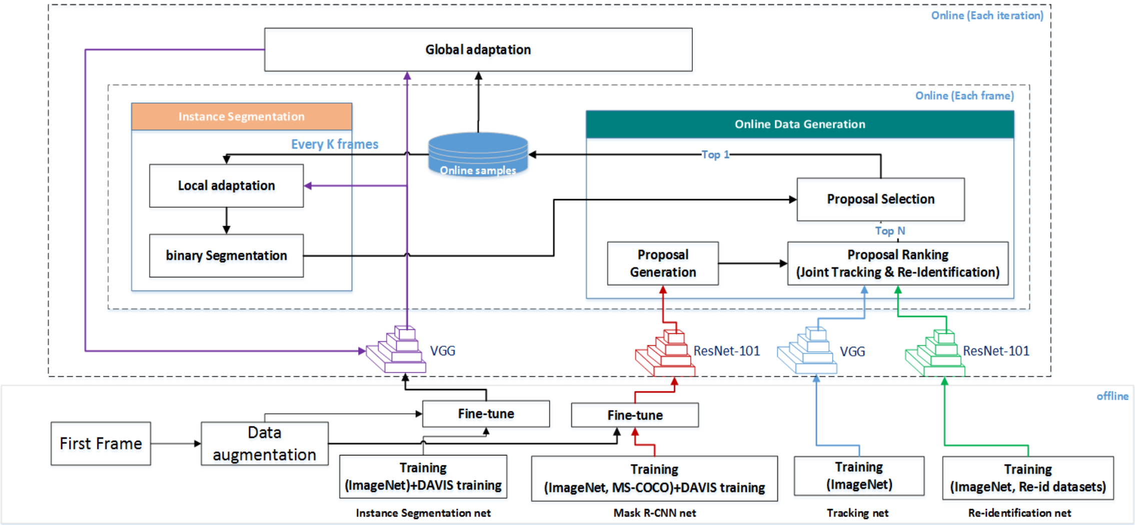


Figure 1. System architecture: The system consists offline training and iterative online instance adaptation.

fine-tuned with DAVIS dataset, including training and augmented data for each testing video. In online training and testing, the proposed adaptive method consists of two modules: online data generation and instance segmentation. The former module aims to generate training data with its mask for each instance in a given frame. The training data will be accumulated across frames and iterations, and be used to update segmentation net in a novel two-level training with different sampling strategy. We will introduce each module in the following subsections.

2.1. Offline Training

In our system, four nets are used for different tasks. In general, all these networks are first trained on the large-scaled and task-specific dataset. In order to adapt to DAVIS domain, detection and segmentation nets are fine-tuned with DAVIS training set. Since each video contains only one frame of annotation, we first augment the annotation to contain variations on both foreground and background and then fine-tune the models to adapt to each video.

2.1.1 Data Augmentation

The lucid data dreaming [5] introduces an elaborate data augmentation method, which synthesizes the foreground changes by rigid and non-rigid transformation with a small extent, and synthesizes the background changes using affine deformations with limited appearance variations. However, due to the movement of both the object and the camera, the background can be totally different across frames. In this paper, external background images were introduced in this paper and stitched with foreground deformations. For each video, we first crawled images from Google. Then we apply Mask R-CNN to each image, and rank the images by the detection numbers in ascending order, and pick top 20 images, indicating that the background is pure. Finally, we apply lucid data dreaming for both first frame and external backgrounds.

2.1.2 Training

Given the above augmentation data, we fine-tune both detection net (Mask R-CNN used in our work) and instance segmentation net. In Mask R-CNN, additional augmentation, including random crop, shrank, mirror, perturbations in hue, contrast, saturation, and brightness are applied with probability 0.5. And only the last layer in RPN and mask branch are fine-tuned, while keeping other layers fixed, in order to keep the network capacity. For the instance segmentation net, we adopt the parent net in the OSVOS [1] which segments each image into foreground and background. Since the DAVIS2018 is a multiple instance dataset, we train our segmentation net by taking all labeled instance as the foreground. Then, in order to adapt the segmentation net to each instance, the net is fine-tuned on the augmented data of each instance respectively.

2.2. Instance Segmentation Module

The instance segmentation module includes binary segmentation nets (the parent net in OSVOS is adopted) for each instance and one fully-connected Conditional Random Field (CRF) [6] classifier. Each segmentation net computes the binary mask for one instance, and the CRF merges all binary masks as the final output. In online training, the segmentation net is updated in two levels: local adaptation and global adaptation. As the video evolves, the local adaptation is conducted every K frames (K=3 in this paper). It fine-tunes the net with high sample rates on data from nearby frames and low rates on those from distant frames,

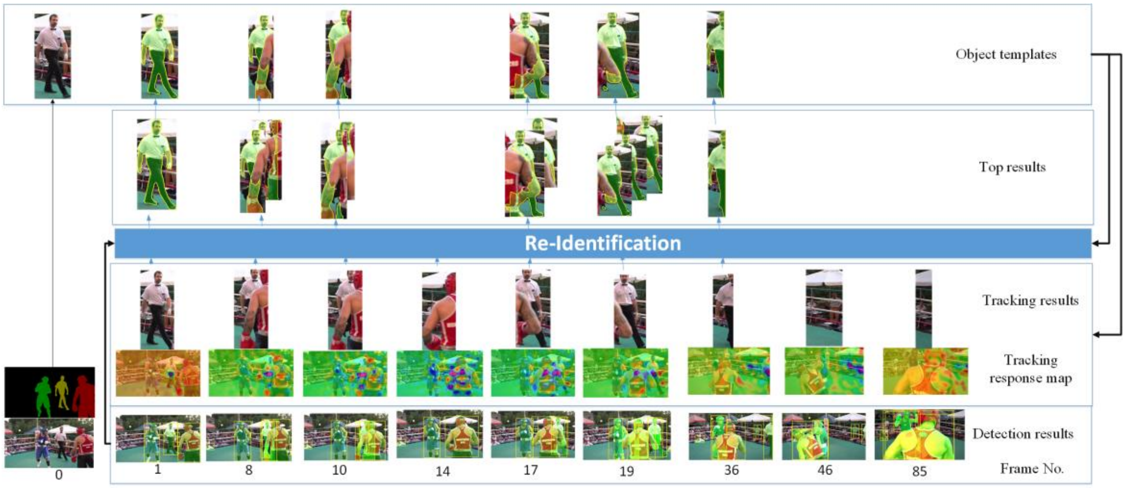


Figure. 2 Illustration of joint tracking and re-identification module. Best view in color.

to gain good representation and good segmentation results on the target frame. After the video evolves, the global adaptation is conducted by fine-tuning the net on all training data with equal sample rates, in order to learn the dynamic variances of the video. The global adaptation and the local adaptation is iteratively conducted, as summarized in the Algorithm 1, where $\mathcal{N}$ is the offline trained weights in section 2.1, $\mathcal{M}_t$ is the binary mask at frame t, $\mathcal{D}$ is the online data set, whose element d_t is the generated data from frame t. Details of the proposal selection will be given in section 2.3.3.

Algorithm 1. Iterative instance segmentation for each instance

```
Input: binary segmentation net $\mathcal{N}$
Output: binary mask set $\{\mathcal{M}_t, t = 1, \dots, T\}$
for $iter$=1…$maximum\_iter$ do
  Initialization: training set $\mathcal{D}$=[], $\mathcal{N}_{iter} = \mathcal{N}$
  for $t$=1…T do
    obtain binary mask: $\mathcal{M}_t \longleftarrow forward(\mathcal{N})$
    generate training data:
        $d_t \longleftarrow proposal\ selection\ with\ th_s$ and $th_J$
    accumulate training set: $\mathcal{D} = \mathcal{D} \cup d_t$
    local adaptation: $\mathcal{N}_{iter} \longleftarrow$ fine-tune $\mathcal{N}_{iter}$ on $\mathcal{D}$
  End for
  global adaptation: $\mathcal{N} \longleftarrow$ fine-tune $\mathcal{N}$ on $\mathcal{D}$
  update proposal selection thresholds $th_s$ and $th_J$
End for
```

In order to reduce outlier noises, we run binary segmentation on a cropped ROI (region of interest), which is an expanded region from the bounding box of the mask in the previous frame. If the shorter edge of the ROI is smaller than 224, we resize it to 224 to help segment small objects.

2.3. Online Data Generation Module

2.3.1 Proposal Generation and Ranking

An input frame is first passed through the Mask R-CNN to obtain proposals for target locations and segmentation masks. As shown in Figure 2, we propose a joint tracking and re-identification method to filter and rank proposals in terms of temporal and appearance similarities.

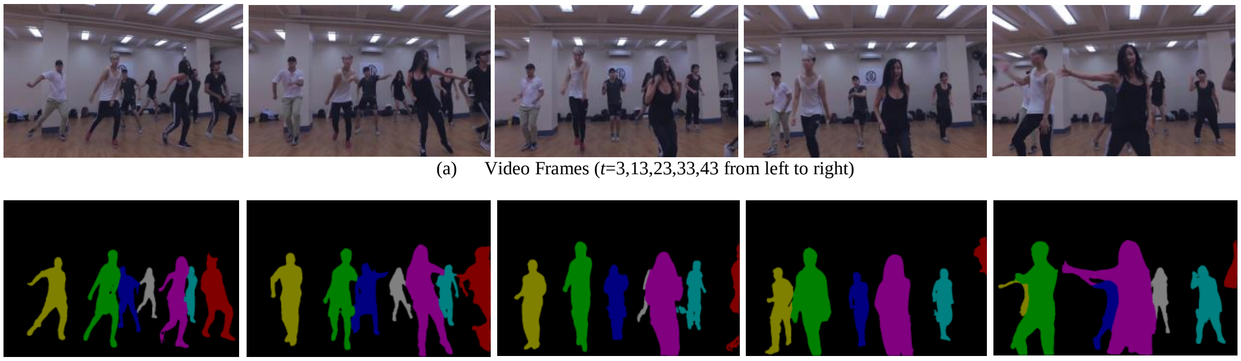
Suppose the t-th input image frame is I_t , the task is to identify the target instance o from a set of detection results $B_t = \{B_{t1}, \dots, B_{tn}\}$. For evaluating appearance similarity, a template set ($\mathcal{S}$) of each object is used which is initialized by the first frame annotation and dynamically augmented by the high confidence bounding box in the rest of frames. To extract appearance features, we extracted global features described in [10]. Then we measure the appearance by calculating average cosine distance between bounding box and templates.

For temporal similarity, we use a deep correlation filter based object tracker to predict possible location and scale in the testing frame [3]. Suppose the tracking bounding box in I_t for o is r_{to}, the final similarity score for matching bounding box B_{tj} with instance o is:

$$s_{jo} = \frac{\alpha}{|\mathcal{S}|} \cdot \sum_{k \in \mathcal{S}} cosine(f_{B_{tj}}, f_k) + \beta \cdot \mathcal{T}(B_{tj}, r_{to}) \quad (1)$$

where $f_.$ indicates the global features extracted from the re-identification convolutional network. The Jaccard $\mathcal{T}$ is used to measure the location closeness between proposal and tracking prediction. $\alpha = 1$ and $\beta = 0.5$ are used to put more weights on appearance similarity. Then we select all proposals which larger than a threshold as possible proposals and passed to the proposal selection module as described in section 2.3.2.

The top-1 proposal is regarded as the best match and added to the template set. If no proposal satisfies the matching threshold, the target is regarded as lost resulted from occlusion or out-of-boundary, as shown in the frame 11 in figure 2. In this case, no sample is added to template set. When the target is lost, we will search for the new appearance in the coming frames with $\beta = 0$ equation (1).



(a) Video Frames (*t*=3,13,23,33,43 from left to right)

(b) Output Instance Masks (*t*=3,13,23,33,43 from left to right)

Figure. 3 Examples of segmentation results.

Finally, to improve the robustness of tracking, we used the available template set to train online tracker, instead of raw tracking result. For frames with no match, the online tracker is not update and stopped, in order to remember the last position and discard wrong samples for training tracker. We observe that this modification can alleviate drift problem during tracking.

2.3.2 Proposal Selection

Given the candidate proposals, the following task is to measure the consistency of each proposal with the binary mask $\mathcal{M}_t$. The final generate sample must meet the following two criteria:

- Color histogram similarity: $s \geq th_s$
- Jaccard similarity: $\mathcal{J} \geq th_J$

It is noted that these two thresholds are reduced during iterations. In early iterations, larger values are assigned. So a small number of samples will be added to $\mathcal{D}$. When relaxing selection condition with smaller thresholds, more training data are generated to learn large appearance variations.

3. Results

We evaluate the proposed method on the DAVIS 2018 challenge dataset. It consists of 150 sequences, with 90 sequences for training, 30 for evaluation and 30 for testing. There are 10, 459 annotated frames and 376 objects. For the testing dataset, each video provides only the first frame annotation. In Figure 3, we show an example of the segmentation results. Our method achieves the fourth place in the semi-supervised DAVIS challenge on video object segmentation 2018 with a region Jaccard of 67.5% and a boundary F measure of 71.5%.

4. Conclusions

In this work, we proposed an adaptive video object segmentation method with online data generation. It is observed that our data generation module can generate high-quality online training samples, which is robust to target deformations, occlusions, and background noise. We also proposed a multi-level training strategy is to fully utilize the training data to improve performance on our segmentation model.

References

[1] Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D. and Van Gool, L., One-shot video object segmentation. In CVPR 2017

[2] Caelles, S., Montes, A., Maninis, K.K., Chen, Y., Van Gool, L., Perazzi, F. and Pont-Tuset, J., The 2018 DAVIS Challenge on Video Object Segmentation. arXiv preprint arXiv:1803.00557, 2018

[3] Danelljan, Martin, Bhat, Goutam, Shahbaz Khan, Fahad and Felsberg, Michael. ECO: Efficient Convolution Operators for Tracking. In CVPR. 2017

[4] He, K., Gkioxari, G., Dollár, P. and Girshick, R., Mask R-CNN. In ICCV, 2017

[5] Khoreva, A., Benenson, R., Ilg, E., Brox, T. and Schiele, B., Lucid Data Dreaming for Object Tracking, The 2017 DAVIS Challenge on Video Object Segmentation. CVPR Workshops, 2017

[6] Krähenbühl, P. and Koltun, V., Efficient inference in fully connected crfs with gaussian edge potentials. In Advances in neural information processing systems. 2011

[7] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., Microsoft coco: Common objects in context. In ECCV 2017

[8] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), pp.211-252, 2015

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

[10] Zhang, Xuan, Luo, Hao, Fan, Xing, Xiang, Weilai, Sun, Yixiao, Xiao, Qiqi, Jiang, Wei, Zhang, Chi and Sun, Jian. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. In ICCV. 2017