The 2018 DAVIS Challenge on Video Object Segmentation - CVPR 2018 Workshops

Class-Agnostic Video Object Segmentation without Semantic Re-Identification

Shuangjie Xu[†] Linchao Bao[‡] Pan Zhou[†]
[†]Huazhong University of Science and Technology [‡]Tencent AI Lab
{shuangjiexu, panzhou}@hust.edu.cn linchaobao@gmail.com

Abstract

This paper presents a class-agnostic video object segmentation approach that won the 3rd place in the 2018 DAVIS Challenge (semi-supervised track). The proposed approach does not use any semantic object re-identification module and thus is more generic to handle unknown types of objects. Specifically, the approach is composed of four steps: 1) An instance proposal box for a given object is predicted from its history trajectories using a linear motion model with explicit occlusion detection; 2) A coarse mask is generated by fusing the warped mask from the preceding frame and the mask prediction from One-Shot Video Object Segmentation (OSVOS) CNN; 3) The coarse mask, truncated by the instance proposal box, is then fed into a mask refinement CNN to get a more detailed mask; 4) An iterative spatio-temporal refinement is lastly performed to get the final segmentation results. For multiple objects case, each single object is dealt with individually and merged into one mask by considering temporal consistency. The effectiveness of the proposed approach is demonstrated with experiments on very challenging sequences.

1. Introduction

Semi-supervised video object segmentation, a labeling task aiming to segment one or more objects from background in a video according to the ground-truth pixels of the given objects in the first frame, has been a crucial task with extensive applications in video editing, video summarization, action recognition, *etc*. Perazzi *et al.* [11] has introduced DAVIS dataset recently, which promotes the development of this field with many deep neural approaches [4, 8, 9] proposed on it. The 2018 DAVIS challenge [5] introduces many pitfalls such as fast motions, small objects, severe inter-object occlusions and so on. To overcome these challenges, we propose a class-agnostic video object segmentation approach with capacity of object recapturing after missing without resorting to semantic re-identification modules.

The pipeline of our approach, shown in Fig 1, consists of four parts: 1) instance bounding box prediction; 2) coarse mask generation; 3) mask refinement CNN; 4) iterative spatio-temporal refinement. In the first part, a linear motion model predict the bounding box of object roughly, together with occlusion detection model to recapture when target lost caused by severe inter-object occlusion or fast motion. Plenty of previous works fix occlusion issue with semantic information such as object detection [9, 10], semantic segmentation [8, 10, 12] and people search [10]. Our occlusion detection model only tasks non-semantic information (optic flow and results of OSVOS [4]) as inputs. Warped instance mask by optic flow [7] along with the OSVOS segmentation result refines the bounding box predicted by motion model. Besides, the OSVOS result is helpful to recapture the lost instance under the temporal knowledge of bounding box history. The third and fourth parts generate instance mask in a cascaded way: the coarse mask is fused by the warped mask and the OSVOS result; the coarse mask and original image croped with the predicted bounding box are then fed into a mask refinement CNN to get mask with more detail. After mask propagation, we adopt the spatio-temporal MRF model [2] to refine generated mask, enforcing the inference result in each frame more like the specific object. With a class-agnostic architecture, our approach is able to segment arbitrary types of objects, and we demonstrate its effectiveness by achieving the 3rd place in the 2018 DAVIS challenge (out of 18 participants).

2. Class-Agnostic Video Object Segmentation

Fig 1 shows the pipeline of proposed approach. The detail of each part is presented in the flowing subsection. As default, the discussion is based on the instance-specific case except the masks margining part in the last subsection.

2.1. Instance Bounding Box Prediction

The bounding box prediction model is based on a linear motion model predicting object position and size with its historical movement. It proved to be that the linear motion model is effective enough except in some extreme cases such as large change in acceleration, serious occlusion and object out of view. Therefore, the occlusion detection model is introduced to recapture the object when the motion model

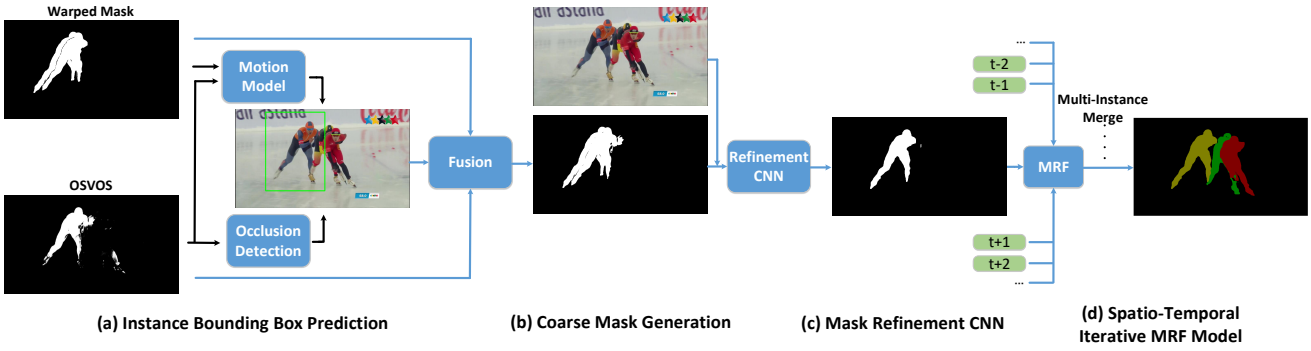

Figure 1. Pipeline of the proposed approach. The MRF model works on the entire sequence.

lost its efficacy.

Linear Motion Model. The bounding box of object in frame F_t depends on two main variable: size $s_t, (w_t, h_t)$ and center point coordinate $p_t, (x_t, y_t)$. In our work, the historical movements in n frame from F_{t-n} to F_{t-1} are adopted as prior knowledge. For the position prediction, the velocity v_t is estimated by

$$v_t = \frac{1}{n}\alpha^{l_t} \sum_{m=t-1}^{t-n} p_m - p_{m-1}, \tag{1}$$

where l_t denotes to the frame number since the last time that object is lost in motion model, and α^{l_t} is an attenuation factor according to the time since object has been lost. Then the predicted center point is obtained by $p_t = p_{t-1} + v_t$. As to the size, average size is taken as the predicted object size $s_t = \frac{1}{n}\beta^{l_t} \sum_{m=t-1}^{t-n} s_m$ with an expansion factor β^{l_t}, since that the change in size is tiny and smooth. With the estimation of p_t and s_t, the motion model gives the bounding box candidate b_t for the following occlusion detection.

Occlusion Detection Model plays an important role in two aspects: 1) refining bbox prediction in the motion model; 2) recapturing object when the motion model lost efficacy. Given the bounding box candidate b_t, we generate a two-dimensional gaussian map G_t with parameters of $\sigma_x^t = {}^{w_t}/_2, \sigma_y^t = {}^{h_t}/_2$, which is used to obtain the weighted mask $W_t = (G_t \cdot O_t > th)$ from OSVOS segmentation result O_t without noise out of the region of interest. The warped mask Q_t from M_{t-1} with optic flow [7, 1, 3] is then merged by W_t to give the coarse mask prediction C_t. The refined bounding box prediction b_t is finnally extracted with the coarse mask C_t by a margin. However, the occlusion detection model will doubt the confidence of b_t if b_t is too small relative to the initial size or very cluttered. In this case, the weighted mask W_t is taken as the recaptured coarse mask with small blobs removed. The recaptured bounding box is then obtained based on W_t.

Algorithm 1 Our Bounding Box Prediction Model.

Input:
original frame F_t, valid warped mask Q_t, OSVOS mask O_t, ground truth mask M_0, center point history $P = \{p_m | m = 0, \ldots, t-1\}$, bbox size history $S = \{s_m | m = 0, \ldots, t-1\}$, number of frames T.

Output:
set recapture flag $R \Leftarrow False$; $l_t \Leftarrow 0$;
for t from 1 to T **do**
  if $\text{sum}(Q_t) < \eta \cdot \text{sum}(M_0)$ **then**
    $R \Leftarrow True$;
  else
    $b_t = \text{linear\_motion\_model}(P, S, l_t)$;
    $C_t = \text{merge\_masks}(Q_t, (G_t \cdot O_t > th))$;
    extract b_t from C_t;
    if b_t is small or b_t changes a lot from history **then**
      $R \Leftarrow True$;
    end if
  end if
  if $R == True$ and object is not stringy or small **then**
    $b_t = \text{linear\_motion\_model}(P, S, l_t)$;
    if $\text{sum}(W_t) < \eta \cdot \text{sum}(M_0)$ **then**
      recapturing failed; $l_t \Leftarrow l_t + 1$;
    else
      extract b_t from W_t; $l_t \Leftarrow 0$; $R \Leftarrow False$;
    end if
  end if
end for
return $\{b_m | m = 1, \ldots, T\}$;

Summarize: The overall flow of the bounding box prediction algorithm is shown in Algorithm 1. The occlusion detection will recapture the bounding box if the b_t is in unreasonable case such as b_t is too small over the initial size, C_t is too clutter or b_t moves fast in comparison to the historical velocity. Recapturing strategy is turned off for objects that are stringy or too small since the OSVOS results for

such objects are not reliable (we use some simple rules to determine if an object is stringy/small based on the shape of its given mask).

2.2. Coarse Mask Generation

The coarse mask C_t is obtained by the fusion of the warped mask Q_t and weighted OSVOS result W_t with an overlap-merging strategy. Define $B_{W_t}=\left\{b^m_{W_t}, m=1,\ldots,K\right\}$ as a set of regions in W_t, the merging strategy can be formulated as:

$$C_t = Q_t \cup \left\{ \bigcup_{m=1}^{K} \phi\left(\text{sum}\left(b^m_{W_t} \cup (Q_t \cap W_t)\right)\right) \cdot b^m_{W_t} \right\}, \tag{2}$$

where function $\phi(x) = \begin{cases} 0, t \leq 0 \\ 1, t > 0 \end{cases}$. The merging strategy takes Q_t as base and merge regions in W_t that have overlapped parts with Q_t into the base.

2.3. Mask Refinement CNN

The mask refinement CNN is built based on the one stream architecture in LucidTracker [8]. In our network, in order to take advantage of multi-scale information, features from pooling layer of block 1 to 4 in vgg architecture are fed to a convolutional layer with 1×1 kernel and an interp layer, and then combined with the features from the ASPP layer in Deeplab v2 [6]. Besides, rather than utilizing full-resolution images, we prefer to take cropped regions (4-channel, RGB+coarse mask) from the images F_t and the coarse masks C_t according to bounding box as network input, which proved to bring more details in the final segmentation. The network is firstly trained offline using training set in 2018 DAVIS; and then trained with Lucid data augmentation [8] online with each challenge sequence. The detailed segmentation is then generated by $S_t = \text{mask\_refinment\_CNN}(F_t, C_t, b_t)$.

2.4. Iterative Spatio-Temporal Refinement

We further refine $S = \{S_m, m = 1,\ldots,T\}$ by adopting a spatio-temporal Markov Random Field (MRF) [2], which is capable of utilizing both spatial and temporal information to refine the masks. With an alternating of spatial mask refinement and temporal fusion, some lost details can be rebuilt in a way of spatio-temporal "voting". Experiments in Sec. 3 show the big improvements with this step. After this step, the fine instance-specific masks $V = \{V_m, m = 1,\ldots,T\}$ are obtained, which is then merged to the final multi-instance segmentation.

The merging strategy is shown in Algorithm 2: spatial confidence firstly decide which object the overlapped blobs belonging to, and the temporal coherency is taken into account as additional information when it is ambiguous to use spatial information only.

Algorithm 2 Our Multi-Instance Merging Strategy.

Input:
fine instance-specific masks $V^i_t, i = 1,\ldots,n$ for all objects, history mask V^i_{t-1}, segmentation probability map from refinement mask CNN $Z^i_t, i = 1,\ldots,n$.
Output:
set multi-instance segmentation Y_t with the object id that has the max value in V_t pixel-by-pixel;
for patch a in all overlap patches **do**
  $Ids \Leftarrow$ all object ids sorted by the value $\text{sum}\left(Z^i_t[a]\right)$ from high to low;
  if $\text{sum}\left(Z_t^{Ids[0]}[a]\right) \cdot \lambda > \text{sum}\left(Z_t^{Ids[1]}[a]\right)$ **then**
    $Y_t[a] \Leftarrow Ids[0]$;
  else
    obtain the warped mask $Q_t^{Ids[0]}$ from $V_{t-1}^{Ids[0]}$, $Q_t^{Ids[1]}$ from $V_{t-1}^{Ids[1]}$;
    if $\text{sum}\left(Q_t^{Ids[0]}[a]\right) > \text{sum}\left(Q_t^{Ids[1]}[a]\right)$ **then**
      $Y_t[a] \Leftarrow Ids[0]$;
    else
      $Y_t[a] \Leftarrow Ids[1]$;
    end if
  end if
end for
return Y_t for the multi-instance segmentation;

3. Experiments

Experimental Setup: in the 2018 DAVIS challenge, we set $\alpha = 0.9$, $\beta = 1.05$, history number $n = 10$ in motion model, $\eta = 0.5$, confidence threshold value $th = 0.3$, $\lambda = 0.8$ in merging strategy. The OSVOS setting follows the work in [4], trained with data augmentation strategy such as random crop, random scale, vertical flip, random changes in brightness, saturation and contrast. For the refinement mask CNN, we first train it online with $500k$ iteration in a lr of $1e^{-3}$, following by another $2k$ iteration online training on $1k$ lucid data for each challenge sequences with $lr = 1e^{-4}$.

Experimental Results: our approach achieve a global mean (scores from Region J and Boundary F) of 69.7 (the 3d place out of 18 participants) on 2018 DAVIS test-challenge set, shown in Table 3 with name "Dawnsix". Some qualitative results are shown in Fig 2.

Ablation Study: due to limited number of submissions in 2018 DAVIS challenge dataset, we investigate the effects of each part in our approach based on 2017 DAVIS test-dev dataset. Taken the linear motion model and coarse mask generation model as our baseline model, we achieve a global mean score 52.8. Shown in Table 3, the occlusion detection model brings an obvious improvement (52.8 $\rightarrow$ 60.1) thanks to its ability to refine bounding box and recapture instance. The iterative spatio-temporal refinement

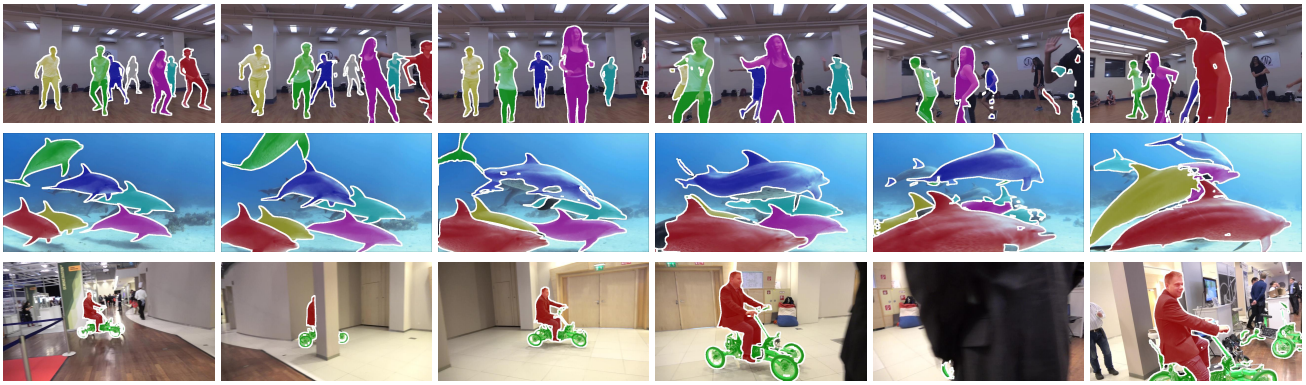

Figure 2. Qualitative results from the 2018 DAVIS challenge. Images are sampled at the average intervals for each sequence.

Table 1. Results on the 2018 DAVIS test-challenge set.

Measure	Jono	Lixx	**Dawnsix**	Team ILC_RIL
Ranking	1	2	**3**	4
Global Mean ↑	74.7	73.8	69.7	69.5
J Mean ↑	71.0	71.9	66.9	67.5
J Recall ↑	79.5	79.4	74.1	77.0
J Decay ↓	19.0	19.8	23.1	15.0
F Mean ↑	78.4	75.8	72.5	71.5
F Recall ↑	86.7	83.0	80.3	82.2
F Decay ↓	20.8	20.3	25.9	18.5

Table 2. Ablation study on the 2018 DAVIS test-dev set.

Approach	score	boost
linear motion model + coarse mask generation model	52.8	-
+ occlusion detection model	60.1	**7.3**
+ mask refinement CNN	62.9	2.8
+ iterative spatio-temporal refinement	**68.1**	5.2

gives another big improvement (62.9 → 68.1) with 10 iterations of TF and MR [2].

4. Conclusion

We have presented a class-agnostic approach for semi-supervised video object segmentation. In contrast to previous works, our approach is able to recapture missing objects utilizing only non-semantic information. The proposed approach does not use any semantic object re-identification module, which is commonly adopted by other participants, and thus is more generic to handle unknown types of objects. Experimental results demonstrate the effectiveness of the proposed approach. In future work, we will replace the occlusion detection model with learned models from data and expect further improvements.

References

[1] L. Bao, H. Jin, B. Kim, and Q. Yang. A comparison of tv-l1 optical flow solvers on gpu. In *GTC Posters*, 2014. 2

[2] L. Bao, B. Wu, and W. Liu. Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In *CVPR*, 2018. 1, 3, 4

[3] L. Bao, Q. Yang, and H. Jin. Fast edge-preserving patchmatch for large displacement optical flow. In *CVPR*, 2014. 2

[4] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*. IEEE, 2017. 1, 3

[5] S. Caelles, A. Montes, K.-K. Maninis, Y. Chen, L. Van Gool, F. Perazzi, and J. Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018. 1

[6] 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. *IEEE PAMI*, 40(4):834–848, 2018. 3

[7] 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*, Jul 2017. 1, 2

[8] 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, 3

[9] X. Li, Y. Qi, Z. Wang, K. Chen, Z. Liu, J. Shi, P. Luo, C. C. Loy, X. Tang, A. Khoreva, et al. Video object segmentation with re-identification. In *CVPR Workshops*, 2017. 1

[10] X.-S. T. Nguyen, 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. 1

[11] 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 *CVPR*, pages 724–732, 2016. 1

[12] 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, 2017. 1