Video Object Segmentation with Re-identification

Xiaoxiao Li[1] Yuankai Qi[2] Zhe Wang[3] Kai Chen[1] Ziwei Liu[1] Jianping Shi[3]
Ping Luo[1] Chen Change Loy[1] Xiaoou Tang[1]

[1]The Chinese University of Hong Kong [2]Harbin Institute of Technology
[3]SenseTime Group Limited

Abstract

Conventional video segmentation methods often rely on temporal continuity to propagate masks. Such assumption suffers from issues like drifting and inability to handle large displacement. To overcome these issues, we formulate an effective mechanism to prevent the target from being lost via adaptive object re-identification. Specifically, our Video Object Segmentation with Re-identification (VS-ReID) model includes a mask propagation module and a ReID module. The former module produces an initial probability map by flow warping while the latter module retrieves missing instances by adaptive matching. With these two modules alternatively updated, our VS-ReID records a global mean (Region Jaccard and Boundary F measure) of 0.699, the best performance in 2017 DAVIS Challenge.

1. Introduction

Video object segmentation in 2017 DAVIS Challenge [13] is non-trivial – a video typically consists of more than one annotated object, with many distractors, small objects and fine structures. The complexity of the problem increases with severe inter-object occlusions and fast motion.

Conventional approaches that rely on temporal continuity suffer from issues like drifting and inability to handle large displacement. To overcome these issues, we formulate an effective mechanism to prevent the target from being lost via adaptive object re-identification. Specifically, our Video Object Segmentation with Re-identification (VS-ReID) model includes a mask propagation module and a ReID module. The mask propagation module is a two-stream convolutional neural network, inspired by [11]. The RGB branch of the mask propagation module accepts a bounding box and a guided probability map as input, and produces a segmentation mask for the main instance in it as output. The guided probability map is obtained from adjacent frames' predictions by flow warping. In addition to the RGB branch, we also train an optical flow branch to incorporate the temporal information. The final segmentation mask of the image patch is obtained by averaging the predictions of these two branches.

To cope with frequent occlusions and large pose variations in dynamic scenes, we leverage object re-identification module to retrieve missing instances. Specifically, when previously missing instances are re-identified with a high confidence, they are assigned with a higher priority to be recovered during the mask propagation process. For each retrieved instance, we take its frame as the starting point and use the mask propagation module to bi-directionally generate the probability maps in its adjacent frames.

With the updated probability maps, the mask propagation module and ReID module of VS-ReID are alternatively applied to the whole video sequence until no more high confidence instances can be found. Finally, for each frame, the instance segmentation results are obtained by merging the probability maps of all the instances. With both flow warping to ensure temporal continuity and object re-identification to recover missing objects, VS-ReID records a global mean (Region Jaccard and Boundary F measure) of 0.699, the best performance in 2017 DAVIS Challenge.

2. Related Work

The realm of object segmentation witnesses drastic progress these days, including the marriage of deep learning and graphical models [17, 9] and the efforts to enable real-time inference on high-res images [7, 16]. Since most of visual sensory data are videos, it is crucial to extend object segmentation from image to video. Existing video segmentation methods [10, 11] rely on temporal continuity to establish spatio-temporal correlation. However, real-life videos observe lots of deformation and occlusion, rendering such assumption suffer from issues like drifting and inability to handle large displacement. In this work, we propose Video Object Segmentation with Re-identification (VS-ReID) to overcome these issues.

3. Approach

Our Video Object Segmentation with Re-identification (VS-ReID) model includes a mask propagation module and a re-identification module. The mask propagation module propagates the probability map from the predicted frames to the adjacent frames. Meanwhile, we employ the re-identification module to retrieve the instances that are missing during the mask propagation. Two modules are alternatively applied to the whole video sequence. Next, we will first present these two modules respectively in Sec. 3.1 and Sec. 3.2, then introduce the algorithm of VS-ReID in Sec. 3.3.

Algorithm 1 Mask propagation for single object

procedure $\mathcal{M}_{mp}(I_i, I_j, P_{i,k})$
  $P_{j,k} \leftarrow 0$ ▷ initialize
  $f_{i\to j} \leftarrow \mathcal{F}(I_i, I_j)$ ▷ extract the optical flow
  $P_{i\to j,k} \leftarrow \mathcal{W}(P_{i,k}, f_{i\to j})$ ▷ flow guided warp
  $b \leftarrow \mathrm{Box}(P_{i\to j,k} > 0.5)$ ▷ obtain the bounding box
  $P^b_{j,k} \leftarrow \mathcal{N}_{mp}(I^b_j, f^b_{i\to j}, P^b_{i\to j,k})$
  return $P_{j,k}$

3.1. Mask Propagation Module

The inference algorithm of mask propagation is summarized in Algorithm 1. Given two adjacent frames I_i, I_j, and the pixel-level probability map for instance k in the frame i, $P_{i,k}$, we aim to predict the probability map for instance k in the frame j, $P_{j,k}$.

Following [11, 6], we first obtain the coarse estimation of $P_{j,k}$, $P_{i\to j,k}$, from $P_{i,k}$ by flow guided warping. We use FlowNet2.0 [5] to extract the optical flow $f_{i\to j}$ between frame i and j. The probability map $P_{i,k}$ is warped to $P_{i\to j,k}$ according to $f_{i\to j}$ by a bilinear warping function $\mathcal{W}$. After that, we employ a convolutional neural network, mask propagation network $\mathcal{N}_{mp}$, to further refine the coarse estimation. Rather than full-images, our mask propagation network accepts size-normalized patches as input and produces the refined probability patch to cope with large scale variations. More specifically, we crop the patches I^b_j, $f^b_{i\to j}$ and $P^b_{i\to j,k}$ from full-image by instance bounding box b. Then we resize those patches into a fixed size and feed them into the mask propagation network to get the probability patch $P^b_{j,k}$. Finally, $P^b_{j,k}$ is resized back to the original size, and fill into a full size zero map to generate the prediction of $P_{j,k}$. Unlike full-image based network [11, 6], our method can easily capture the small objects and fine structures.

Mask Propagation Network. As shown in Fig. 1, our mask propagation network is a two-stream convolutional neural network, inspired by [6]. However, several important modifications are necessary to further improve the network performance. First, we adopt the much deeper ResNet-101 [4] network to increase the model capacity. Second, as we mentioned before, since our network take the patches as input, it is capable of capturing more details compared with full-image based network. We also slightly enlarge the bounding box to keep more contextual information. Third, to increase the resolution of prediction, we enlarge the size of feature maps by decreasing the convolutional stride and replace convolutions by the dilated convolutions. Similar to [1], atrous spatial pyramid pooling and multi-scale testing are also employed. Last but not least, after independently training, two streams are jointly fine-tuned which further improves the performance.

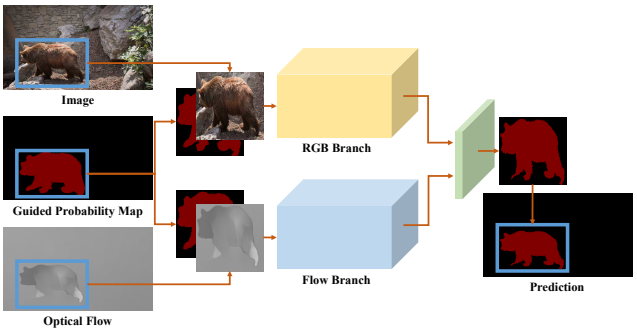

Figure 1. Network architecture of mask propagation network. **Best viewed in color.**

Algorithm 2 Re-identification module

procedure $\mathcal{M}_{reid}(I_i, P_{i,k}, t_k)$
  $X \leftarrow \mathcal{N}_{det}(I_i)$ ▷ obtain the candidate boxes
  for $x_j \in X$ **do**
    $s_j \leftarrow S_C(\mathcal{N}_{reid}(I_i^{x_j}), \mathcal{N}_{reid}(t_k))$ ▷ S_C denotes the cosine similarity
  $\hat{j} \leftarrow \arg\max_{j \le |X|} s_j$
  $b \leftarrow \mathrm{Box}(P_{i,k} > 0.5)$
  if $s_{\hat{j}} > \rho_{reid}$ **and** $\mathrm{IoU}(x_{\hat{j}}, b) < \rho_{occ}$ **then**
    return $x_{\hat{j}}, s_{\hat{j}}$
  else
    return $x_{\hat{j}}, -1$ ▷ fail or unnecessary

3.2. Re-identification Module

Our mask propagation module is based on the short-term memory and highly rely on temporal continuity. However, frequent occlusions and large pose variations are very common in dynamic scenes and likely to cause the propagation failed. To overcome these issues, we leverage object re-identification module to retrieve missing instances. Re-identification module incorporates the long-term memory, which complements with mask propagation module and makes our system more robust.

As summarized in Algorithm 2, during the alternatively inference in VS-ReID, our re-identification module takes an single frame I_i, the current pixel-level probability map $P_{i,k}$ which is predicted in the previous round of inference

Figure 2. Pipeline of our Video Object Segmentation with Re-identification (VS-ReID) model. **Best viewed in color.**

for instance k in the frame i, and the template of instance k, t_k as input, produces the retrieved boundary box x, and corresponding re-identification score s. In this module, we obtain the candidate bounding boxes X in frame I_i through a detection network $\mathcal{N}_{det}$. For each candidate bounding box x_j, the re-identification score between x_j and t_k is conducted through measuring the cosine similarity between their features that are extracted from a re-identification network $\mathcal{N}_{reid}$. Suppose $x_{\hat{j}}$ is the most similar candidate bounding box, it is only accepted as the final result if two conditions are satisfied: First, $x_{\hat{j}}$ is sufficiently similar with the template t_k, that is, the re-identification score between $x_{\hat{j}}$ and t_k is larger than a threshold ρ_{reid}; Second, current $P_{i,k}$ is not consistent with $x_{\hat{j}}$, otherwise we don't need to retrieve the instance k in frame i. To evaluate this condition, we compute the IoU score between $x_{\hat{j}}$ and current bounding box from $P_{i,k}$. If it is less than another threshold ρ_{occ}, which means they are inconsistent, we believe that we have made a wrong prediction of $P_{i,k}$ in the previous rounds and accept $x_{\hat{j}}$ as the retrieve bounding box. Those two thresholds are selected on the validation set.

Detection & Re-identification Network. We directly adopt the Faster R-CNN [14] as our detection network $\mathcal{N}_{det}$. For the re-identification network $\mathcal{N}_{reid}$, we employ the architecture of 'Identification Net' in [15] and retrain this network to adapt the general object re-identification task.

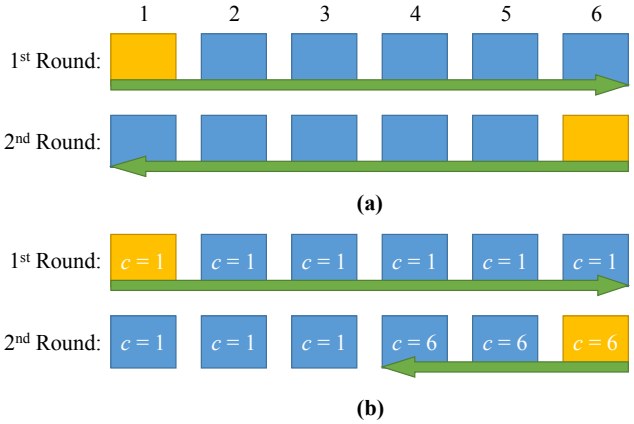

Figure 3. Existing probability maps might be impaired during the alternatively updating. Therefore, we devise a checkpoint mechanism to avoid this issue. **Best viewed in color.**

3.3. VS-ReID

In this section, we will introduce the VS-ReID algorithm that combine previous two components to inference the masks of all instances on the whole video sequence.

As shown in Fig. 2, given a video sequence and the mask (*i.e.* ground-truth probability map) of the objects in the

Algorithm 3 VS-ReID algorithm

procedure VS-REID($\{I\}, \{P_1\}$)
  $N \leftarrow |\{I\}|$ ▷ number of frames
  $K \leftarrow |\{P_1\}|$ ▷ number of instances
  for k = 1 **to** K **do**
    obtain the template t_k from $P_{1,k}$
  for i = 2 **to** N **do** ▷ initialize probability maps
    for k = 1 **to** K **do**
      $P_{i,k} \leftarrow \mathcal{M}_{mp}(I_{i-1}, I_i, P_{i-1,k})$
      $c_{i,k} \leftarrow 1$
  loop
    $\hat{s} \leftarrow -1$
    for i = 2 **to** N **do** ▷ retrieve instances
      for k = 1 **to** K **do**
        $x, s \leftarrow \mathcal{M}_{reid}(I_i, P_{i,k}, t_k)$
        if $s > \hat{s}$ **and** $c_{i,k} \neq i$ **then**
          $\hat{s} \leftarrow s, \hat{x} \leftarrow x, \hat{i} \leftarrow i, \hat{k} \leftarrow k$
    if $\hat{s} < 0$ **then**
      break ▷ no instance retrieved
    else
      $P_{\hat{i},\hat{k}} \leftarrow 0, f_{\hat{i}} \leftarrow \mathcal{F}(I_{\hat{i}}, I_{\hat{i}+1})$
      $b \leftarrow \mathrm{Box}(P_{1,\hat{k}} > 0.5)$
      $P^{\hat{x}}_{\hat{i},\hat{k}} \leftarrow \mathcal{N}_{mp}(I^{\hat{x}}_{\hat{i}}, f^{\hat{x}}_{\hat{i}}, P^{b}_{1,\hat{k}})$ ▷ recover
      for $i = \hat{i} + 1$ **to** N **do** ▷ forward propagate
        if $|c_{i,\hat{k}} - i| > |\hat{i} - i|$ **then**
          $P_{i,\hat{k}} \leftarrow \mathcal{M}_{mp}(I_{i-1}, I_i, P_{i-1,\hat{k}})$
          $c_{i,\hat{k}} \leftarrow \hat{i}$
      for $i = \hat{i} - 1$ **downto** 2 **do** ▷ backward propagate
        if $|c_{i,\hat{k}} - i| > |\hat{i} - i|$ **then**
          $P_{i,\hat{k}} \leftarrow \mathcal{M}_{mp}(I_{i+1}, I_i, P_{i+1,\hat{k}})$
          $c_{i,\hat{k}} \leftarrow \hat{i}$
  return $\{P\}$

first frame, VS-ReID first initializes the probability maps $\{P\}$. We enumerate all instances and forward propagate their probability maps from the first frame to the last frame by mask propagation module. After initialization, the re-identification module and mask propagation module are alternatively applied to the whole video sequence until no more high confidence instances can be found. To be more specific, we first applied re-identification module to the whole video for all instances. We keep the retrieved bounding box $\hat{x}$ with the highest similarity score $\hat{s}$. Suppose $\hat{x}$ is the bounding box of instance $\hat{k}$ in frame $\hat{i}$, we then try to recover the probability map of instance $\hat{k}$ in frame $\hat{i}$, $P_{\hat{i},\hat{k}}$. The recover process is quite similar with mask propagation module, with one difference: there is no guided probability map from adjacent frames. So we replace that with the probability patch of instance $\hat{k}$ cropped from the first frame. Once we obtain the recovered probability map, we can take it as the starting point and use the mask propagation module to bi-directionally recover more probability maps of instance $\hat{k}$ in adjacent frames. However, sometimes existing probability maps will be impaired during this alternatively updating. An example is shown in Fig. 3 (a), suppose we have 6 frames in a video sequence. In the first round of alternatively updating, we retrieve the instance k in the first frame and propagate the recovered mask to the end of video sequence. In the second round, we retrieve the instance k again in the last frame and do the backward propagation. In this case, all probability maps we predicted in the first round will be overwritten. Because of the longer propagation distance, the probability map for instance k in the second frame might be impaired in the second round. To avoid this issue, we devise a checkpoint mechanism with a new variable $c_{i,k}$ recording the starting point by which $P_{i,k}$ is updated. The initial value of $c_{i,k}$ is 1, and every probability map prefers to be updated by a closer starting point. As shown in Fig. 3 (b), the backward propagation will be interrupted at the fourth frame, since the first frame is closer to the third frame compare with the last one. Finally, we combine all $\{P\}$ to generate the mask prediction M through:

$$M_i(l) = \arg \max_{0 \leq k \leq K} \frac{1}{Z} * \begin{cases} P_{i,k}(l) & k \neq 0 \\ \prod_{j=1}^{K}(1 - P_{i,j}(l)) & k = 0 \end{cases}$$

where $Z = \prod_{j=1}^{K}(1 - P_{i,j}(l)) + \sum_{j=1}^{K} P_{i,j}(l)$ is a normalizing factor, i is the frame index, l is a pixel's location, K is the number of instances in the video sequence.

3.4. Implementation Details

Two branches of mask propagation network are first trained individually. The RGB branch is pre-trained on the MS-COCO [8] and PASCAL VOC [3] dataset. During the pre-training, We use the deform ground-truth mask as the guided probability map. After that, network is fine-tuned on the DAVIS training set. The flow branch is initialized by RGB branch's weights and fine-tuned on the DAVIS training set. Finally, those two branches are joint fine-tuned together on the DAVIS training and validation sets.

Detection and re-identification network are trained on the ImageNet [2] dataset, we followed the training strategy in original papers [14, 15]. In particular, for the person category, we directly use the network in [15] as our re-identification network.

4. Experiments

We evaluate our VS-ReID on the DAVIS 2017 [13] datasets. DAVIS 2017 dataset contains 150 video sequences with all frames annotated with hight-quality object masks. There are 60 videos in the *train set*, 30 videos in the *val set*, 30 videos in the *test-dev set* and 30 videos in the *test-challenge set*. In our experiments, we employ both *train set* and *val set* for training, and all performances are reported on the *test-dev set*. Followed [12], we adopt region($\mathcal{J}$) and boundary($\mathcal{F}$) measures to evaluate the performance.

Table 1. Ablation study of each module in VS-ReID.

	$\mathcal{J}$-mean	$\mathcal{F}$-mean	global-mean	boost
baseline[11]	0.509	0.526	0.517	-
+ full-image to bbox	0.532	0.577	0.555	+ 0.038
+ flow-stream	0.568	0.600	0.584	+ 0.007
+ re-id module	0.633	0.670	0.652	+ 0.068
+ multi-scale testing	0.644	0.678	**0.661**	+ 0.009

Figure 4. Missing instances are retrieved back by re-identification module. We annotate the retrieved instances by blue bounding boxes. **Best viewed in color.**

4.1. Ablation Study

In this section, we investigate the effects of each component in VS-ReID model. Table 1 summarizes how performance gets improved by adding each component step-by-step into our VS-ReID model.

We choose [11] as our baseline model. After modified the input from full-image to bounding box, global-mean increases 3.8 percent and the boundary ($\mathcal{F}$) measure achieves significant improvement of 5.1 percent. It demonstrates that bounding box input overcomes the large scale variations and contributes to capture the boundary details. As mentioned in Sec. 3.1, to incorporate the temporal information, we train an optical flow branch and joint fine-turning it with the RGB branch. This two-stream architecture also slightly improve the performance. Employing the alternative algorithm we introduced in Sec. 3.3 greatly improves the global-mean by 6.8 percent, which shows that the re-identification module and alternative algorithm are essential. We also visualize the example videos where are improved by this alternative algorithm in Fig. 5. Once an instance is recovered, it will be also benefit to adjacent frames' prediction. Finally, multi-scale testing further improves the results.

4.2. Benchmark

As shown in Table 2, VS-ReID achieves a global mean of 0.699 on *test-challenge set*, the best performance in 2017 DAVIS Challenge. By inspecting closer, we observe that VS-ReID wins both $\mathcal{J}$-Mean and $\mathcal{F}$-Mean measures and outperforms the second place method more than 2 percent. Thanks to the re-identification module that incorporates the long-term memory, our $\mathcal{J}$-Decay and $\mathcal{F}$-Decay are also relatively small. In Fig. 5, we demonstrate some examples of VS-ReID prediction on DAVIS *test-dev set* and *test-challenge set*.

5. Conclusion

In this work we tackle the problem of video object segmentation and explore the utility of object re-identification. We propose Video Object Segmentation with Re-identification (VS-ReID) model which includes two dedicated modules: a mask propagation module and a ReID module. It is observed that our ReID module combined with bidirectional refinement is capable to retrieve missing instances and greatly improve the performance. These two modules are alternatively updated and jointly trained, enabling our final model to win the DAVIS video segmentation challenge.

References

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 2

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[3] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2), 2010. 4

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 2

[5] 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*, 2017. 2

[6] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking, 2017. 2

[7] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *CVPR*, 2017. 1

[8] 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. 4

[9] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *ICCV*, 2015. 1

[10] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Deep learning markov random field for semantic segmentation. *arXiv preprint arXiv:1606.07230*, 2016. 1

[11] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A.Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, 2017. 1, 2, 5

Table 2. Results on 2017 DAVIS Challenge *test-challenge set*.

Measure	Ours	Apata	Vanta	Haamo	Voigt	Lalal	Cjc	YXLKJ	Wasid	Froma	Zwrq0	Drbea	Anews	Ilanv	Koh	Make	Kozab	Xn881	Zpd	Griff	Nitin	Team5
Ranking	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Global Mean↑	**69.9**	67.8	63.8	61.5	57.7	56.9	56.9	55.8	54.8	53.9	53.6	51.9	50.9	49.7	49.1	48.0	47.8	47.6	47.1	42.0	25.6	11.2
$\mathcal{J}$ Mean↑	**67.9**	65.1	61.5	59.8	54.8	54.8	53.6	53.8	51.6	50.7	50.5	50.5	49.0	46.0	45.9	46.3	43.9	47.8	44.9	40.6	24.9	11.8
$\mathcal{J}$ Recall↑	**74.6**	72.5	68.6	71.0	60.8	60.7	59.5	60.1	56.3	54.9	54.9	56.4	55.1	49.3	50.2	50.0	45.8	56.3	48.0	42.1	12.3	7.3
$\mathcal{J}$ Decay ↓	22.5	27.7	17.1	21.9	31.0	34.4	25.3	37.7	26.8	32.5	28.0	34.1	21.3	33.1	36.1	40.2	33.0	16.7	31.8	37.4	13.1	14.0
$\mathcal{F}$ Mean↑	**71.9**	70.6	66.2	63.2	60.5	59.1	60.2	57.8	57.9	57.1	56.7	53.3	52.8	53.3	52.3	49.7	51.6	47.3	49.3	43.3	26.3	10.6
$\mathcal{F}$ Recall↑	79.1	**79.8**	79.0	74.6	67.2	66.7	67.9	62.1	64.8	63.2	63.5	57.9	58.3	58.4	57.1	52.8	56.0	53.0	54.4	43.2	9.1	3.0
$\mathcal{F}$ Decay↓	24.1	30.2	17.6	23.7	34.7	36.1	27.6	42.9	28.8	33.7	30.4	39.5	23.7	36.4	39.2	44.8	36.3	21.6	36.2	40.2	13.0	12.6

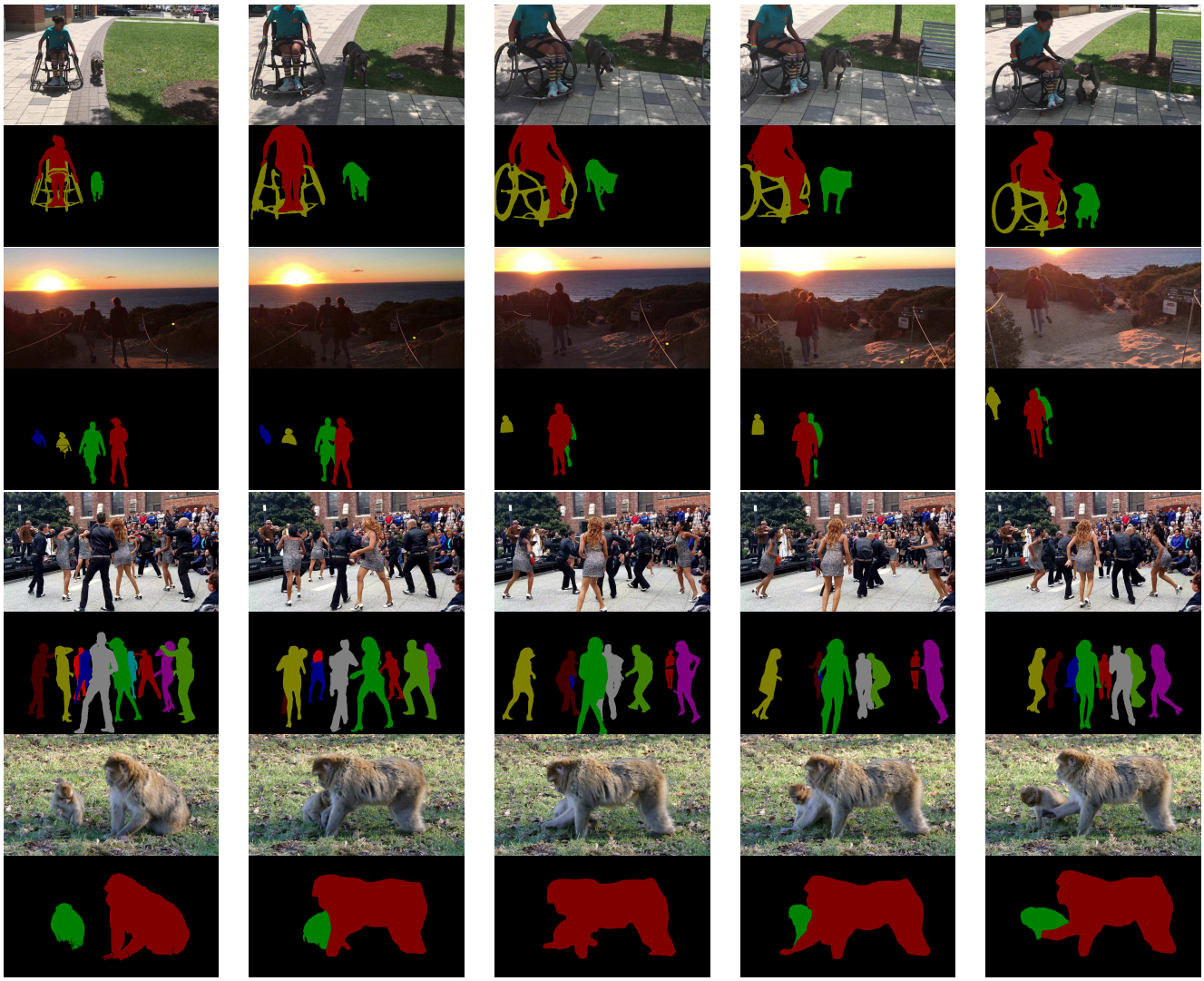

Figure 5. Qualitative results of our VS-ReID model on DAVIS 2017 *test-dev set* and *test-challenge set*.

[12] 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*, 2016. 4

[13] 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, 4

[14] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 3, 4

[15] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017. 3, 4

[16] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. Icnet for real-time semantic segmentation on high-resolution images. *arXiv preprint arXiv:1704.08545*, 2017. 1

[17] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 1