Instance Re-Identification Flow for Video Object Segmentation

Trung-Nghia Le [*1], Khac-Tuan Nguyen[2], Manh-Hung Nguyen-Phan[2], That-Vinh Ton[2], Toan-Anh Nguyen[2], Xuan-Son Trinh[2], Quang-Hieu Dinh[2], Vinh-Tiep Nguyen[3], Anh-Duc Duong[3], Akihiro Sugimoto[4], Tam V. Nguyen[5], and Minh-Triet Tran[2]

[1]The Graduate University for Advanced Studies (SOKENDAI), Japan
[2]University of Science, VNU-HCM, Vietnam
[3]University of Information Technology, VNU-HCM, Vietnam
[4]National Institute of Informatics, Japan
[5]University of Dayton, US

Abstract

In this work, we propose an Instance Re-Identification Flow (IRIF) for video object segmentation. For the instance re-identification task, we focus on two main categories: human and non-human object instances. We track each instance and detect it when it re-appears to determine its corresponding bounding box in video frames. When a non-human object re-appears, we use a list of recent SVM classifiers to segment that object. Otherwise, we use Pyramid Scene Parsing (PSP) Network to automatically segment that person as an initial mask to continue mask propagation. In particular, we use object detector, Faster R-CNN, to detect person and extract person attribute as a key feature for both tracking and re-identification. In addition, DeepFlow and Deformable Part Model (DPM) are utilized to track and detect non-human objects. Regarding object segmentation, we adopt multi-SVM classifiers embedding history reference with several unary components, namely, saliency, CNN features, location and color, to segment each object instance within its possible bounding box in each frame. Note that we also estimate the z-order of each instance to enhance the later instance tracking and mask propagation. Boundary snapping is adopted to further refine instance shapes. Finally, our IRIF method achieves very competitive results in DAVIS Challenge 2017, namely, 0.615, 0.662, and 0.638 in terms of region similarity (Jaccard index), contour accuracy (F-measure), and global score, respectively.

[*]Corresponding author. Email: ltnghia@nii.ac.jp

1. Introduction

Video object segmentation is considered as a labeling problem aiming to separate foreground instance object(s) from the background region of a video. Object segmentation in videos is very beneficial in a wide range of practical applications, *i.e.*, action recognition, object tracking, video summarization, scene understanding, autonomous vehicle, and surveillance system.

Regarding the related works, there exist two main streams of approaches to solve this interesting problem. The first type of methods is based on bottom-up approaches. These such methods extend the concept of salient object detection to videos[8, 9, 12, 13, 11]. Some of these methods generate several ranked segmentation hypotheses [10]. The main advantage is that they do not require any manual annotation and do not assume any prior information on the object to be segmented. In addition, they are well suited for parsing large scale databases. However, these bottom up methods do not exploit the history information of the objects. Therefore, they may have problems when an object re-appears after missing in few video frames. Also, they are prone to errors when there exist many salient objects with non-tracking requirement in the videos.

The second type of methods is proposed in a top-down manner. Tsai *et al.* [18] propose an efficient object flow algorithm that considers video segmentation and optical flow estimation simultaneously. For the segmentation model, they construct a multi-level graphical model that consists of pixels and superpixels, each of which plays different roles for segmentation. At the superpixel level, each superpixel is likely to contain pixels from the foreground and background as the object boundary may not be clear. At the

pixel level, each pixel is less informative although it can be used for more accurate estimation of motion and segmentation. With the combination of these two levels, the details around the object boundary can be better identified by exploiting both statistics contained in superpixels and details in the pixel level. Meanwhile, Jampani *et al.* [5] introduce Video Propagation Networks (VPN) that propagate information forward through video data. The VPN architecture is composed of two components. A temporal bilateral network that performs image adaptive spatio-temporal dense filtering. The bilateral network allows to connect densely all pixels from current and previous frames and to propagate associated pixel information to the current frame. The bilateral network allows the specification of a metric between video pixels and allows a straight-forward integration of temporal information. This is followed by a standard spatial CNN on the bilateral network output to refine and predict for the present video frame. In deep learning research, Convolutional Neural Networks (CNNs) have shown outstanding performance in many fundamental areas in computer vision, enabled by the availability of large-scale annotated datasets (*e.g.*, ImageNet classification [7]). By treating video object segmentation as a guided instance segmentation problem, Khoreva *et al.* [6] propose to use a pixel labelling CNNs for frame-by-frame segmentation. In particular, given a rough mask estimate from the previous frame $t-1$, they train a CNN to provide a refined mask output for the current frame t. Recently, Caelles *et al.* [1] propose One-Shot Video Object Segmentation (OSVOS), based on a fully-convolutional neural network architecture that is able to successively transfer generic semantic information, learned on the large scale ImageNet dataset, to the task of foreground segmentation, and finally to learning the appearance of a single annotated object of the test sequence (so-called one-shot learning).

In literature, there exist many efforts on constructing datasets towards this interesting problem. Firstly, the Densely-Annotated VIdeo Segmentation (DAVIS) initiative [14] provided a dataset with 50 high-definition sequences with all their frames annotated with object masks at pixel-level accuracy. Later, a new dataset [15] is constructed for 2017 DAVIS Challenge on Video Object Segmentation in order to further push the performance in video object segmentation. The new dataset consists of a new, larger, more challenging videos. As the main new challenge, the new sequences have more than one annotated object in the scene. In addition, the complexity of the videos has also increased with more distractors, smaller objects and fine structures, more occlusions and fast motion, etc. Figure 1 highlights the difference of DAVIS 2017 dataset with its previous version.

In order to tackle this challenging and interesting problem, we propose an Instance Re-Identification Flow (IRIF) for video semantic segmentation. Our results on DAVIS 2017 Challenge dataset highly indicate that our method is competitive among the state-of-the-art methods in this newly built dataset.

Figure 1. The major difference between annotations of DAVIS 2016 (left) DAVIS 2017 (right) datasets: bi-instance segmentation vs. multi-instance segmentation.

2. Proposed Method

In this section, we introduce our approach of Instance Re-Identification Flow (IRIF) for video object segmentation. Figure 2 illustrates the flow chart of the proposed framework.

2.1. Instance Re-Identification for Video Object Segmentation

Given the first frame with its ground truth label, we extract the bounding box for each instance. Then, we extract CNN features from each bounding box and perform human classification in order to identify the human instance needed to be segmented in the video. In this work, we consider two types of instances: human and non-human ones. For each video frame, we localize the instances in a re-identification manner. For human instances, we first detect person from the image by using Faster R-CNN [16]. Then, we extract person re-identification feature [19] for all detected person region. For the implementation, we use the state-of-the-art implementation to detect[1] and extract person re-identification feature[2]. DeepFlow and Deformable Part Models (DPM) [4] are utilized to track and detect non-human objects. Note that we expand the bounding box to 10% in order to well capture the whole area of the object instances.

Regarding the instance segmentation task, we utilize multiple binary SVM classifiers [2] which is learned from the appearance of the previous n frames with sampling

[1] https://github.com/Eniac-Xie/faster-rcnn-resnet
[2] https://github.com/ShuangLI59/person_search

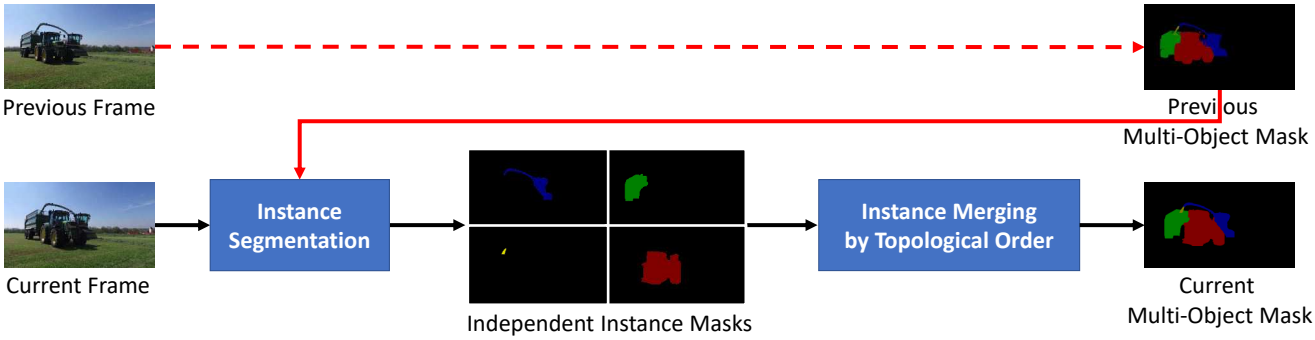

Figure 2. The overview of our proposed IRIF framework. The segmentation performed on the current frame is based on the history information of the previous frames. The merging is enhanced with z-order. The segmentation result of the current frame is further fed to the process of the coming frame.

step size δ, where n and δ are set as 8 and 2, respectively. Note that our multiple binary SVM classifiers are implemented for history reference with several unary instances, *e.g.*, saliency [11], CNN features [7], location of the bounding box and color, to segment each object within its tracked bounding box in each frame. We only update the SVM model if the size of one object instance significantly changes. In case the human is missing and reappear in the next couple of frames, we adopt the state-of-the-art image parser, Pyramid Scene Parsing (PSPNet) [21] with the pre-trained model on PASCAL VOC dataset [3]. The re-identification results from PSPNet are blended into our final segmentation outcomes. We utilize GrabCut [17] for each instance in order to separate it with the background. After this step, each pixel is assigned with the instance ID. Then, we simultaneously run mask propagation for each object instance as described in the following subsection.

2.2. Instance Topological Order Estimation for Mask Propagation

Actually, Object Flow [18] can be used for mask propagation with a single instance. However, in case multiple instances to be segmented from a video clip, it is essential to determine the topology relationship (in term of z-order) between components so that we can sequentially combine the propagated masks of different instances into final result.

Let A and B be two instances of interest, M_A and M_B be their initial masks, respectively. We define $A > B$ if A is likely to be closer to the camera than B. From the initial masks of all instances, we aim to detect all partial z-order relationships between pairs of instances by the following heuristics:

- **Rule 1-Connected Components**: If M_B is divided into multiple connected instances by M_A, then $A > B$. In Figure 3a, the green bird is closer to camera than the red one is the green mask splits the red mask into two connected instances.

- **Rule 2-Instance Size**: If M_A does not contains any subregion of M_B (and vice-versa), and we cannot infer their z-order by other topological relationship, we define $A > B$ if the size of M_A is greater than that of M_B. This heuristic is from the observation that a closer object tends to be larger than the other. In Figure 3b, the red object has larger size and is predicted to be closer to camera than the green one. This heuristic is shown to be efficient with *choreography* to estimate the topology order of persons from their mask sizes (Figure 3c).

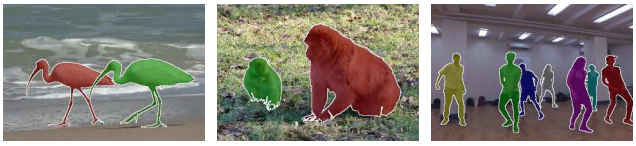

Figure 3. z-order estimation from masks with connected components rule (a), instance size (b), and combining the two rules (c).

Note that we first compute the z-order of different instance masks right after obtaining the ground truth of the first frame. Currently we only apply these two simple rules for this step to illustrate our key idea: use z-order to sequentially combine propagated masks of different instances. More complicated rules will be added into our framework to fully process various practical scenarios. We also update z-order for every single video frame.

2.3. Segmentation Refinement with Boundary Snapping and Rare Instance Attention

Through preliminary results, we observe that the initial segmentation are not smooth enough. Therefore, to improve

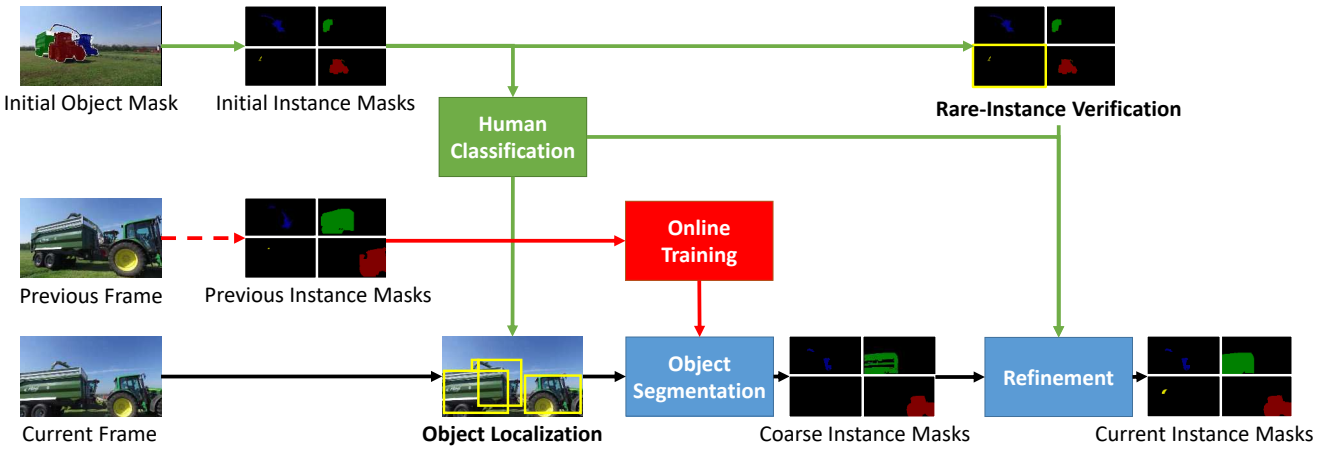

Figure 4. The instance segmentation process in our IRIF framework. Basically, the process involves the instance re-identification with further refinement via rare instance attention and boundary snapping.

the quality of segmentation, we also adopt boundary snapping [1] to further refine object shapes. To this end, we extract the saliency [11] and the contour [20] from the video frame. The salient pixels close to the contour are snapped.

Furthermore, we refine the results by taking the rare instances into consideration. We observe that rare objects are shrunk due to larger objects. To identify rare object instances, we compute the percentage of each object instance mask in terms of area (provided in the first frame). Instances with their size smaller than 5% the total size of tracking objects are considered as the rare ones. Our assumption is that a smaller object tends to be small in the whole video. Next, we recover rare object instances by transferring the results produced by the foreground probability obtained from the binary-SVM classifier on each object instance. Figure 4 illustrates the segmentation process in our proposed IRIF framework.

3. Evaluation on 2017 DAVIS Challenge Dataset

As aforementioned, the main new challenge added to the sequences of DAVIS 2017 Challenge is the presence of multiple objects (instances) in a video frame. Overall, the DAVIS 2017 dataset consists of 150 sequences, totaling $10,459$ annotated frames and 376 objects. There are total 30 video sequences for testing and their ground truth not publicly available. Submissions to all phases is done through the CodaLab site of the challenge[3]. For the evaluation metrics, the per-object measures are used as described in [14]: Region Jaccard (J) and Boundary F measure (F). The overall measures are computed as the mean between J and F, both averaged over all objects. As shown in Table 1, our IRIF method achieves very promising results in DAVIS Challenge 2017, namely, 0.615, 0.662, and 0.638 in terms of region similarity (Jaccard index), contour accuracy (F-measure), and global score, respectively. Our results on DAVIS 2017 Challenge dataset highly indicate that our method is competitive among the state-of-the-art methods in this newly built dataset. Furthermore, Figure 5 visualizes our video object segmentation results. From left to right, we can observe the first video frame, and two pairs of processed video frames (without and with refinement, respectively). Our final results successfully track and segment the key instances. Regarding the runtime performance, it approximately takes around 30 seconds to process one video frame in average.

[3]https://competitions.codalab.org/competitions/16526

4. Conclusion and Future Works

In this paper, we introduce the novel IRIF framework, instance re-identification flow, for semantic segmentation in videos. Our framework is able to segment multiple object instances unlike the binary labeling problem stated in the related works. Throughout the experiments, our framework achieves a competitive result among the participating submissions.

In the future, we are looking at embedding more advanced techniques for different parts in our IRIF framework to improve both quality and runtime performance. In addition, we consider modeling the semantic relationship among object instances into the segmentation process.

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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Table 1. The performance of different methods in DAVIS 2017 Challenge. The rankings in each categories are placed in parentheses. Our results are marked in **blue**.

#	Team	Global	Region J			Boundary F		
		Mean ⇑	Mean ⇑	Recall ⇑	Decay ⇓	Mean ⇑	Recall ⇑	Decay ⇓
1	Lixx	0.699 (1)	0.679 (1)	0.746 (1)	0.225 (7)	0.719 (1)	0.791 (2)	0.241 (6)
2	Apata	0.678 (2)	0.651 (2)	0.725 (2)	0.277 (10)	0.706 (2)	0.798 (1)	0.302 (9)
3	**Vanta**	**0.638 (3)**	**0.615 (3)**	**0.686 (4)**	**0.171 (4)**	**0.662 (3)**	**0.790 (3)**	**0.176 (2)**
4	Haamo	0.615 (4)	0.598 (4)	0.710 (4)	0.219 (6)	0.632 (4)	0.746 (4)	0.237 (5)
5	Voigt	0.577 (5)	0.548 (5)	0.608 (5)	0.310 (22)	0.605 (5)	0.672 (6)	0.347 (12)
6	Lalal	0.569 (6)	0.548 (5)	0.607 (6)	0.344 (18)	0.591 (7)	0.667 (7)	0.361 (13)
7	Cjc	0.569 (6)	0.536 (7)	0.595 (8)	0.253 (8)	0.602 (6)	0.679 (5)	0.276 (7)
8	YXLKJ	0.558 (7)	0.538 (6)	0.601 (7)	0.377 (21)	0.578 (9)	0.621 (11)	0.429 (20)
9	Wasid	0.548 (8)	0.516 (8)	0.563 (10)	0.268 (9)	0.579 (8)	0.648 (8)	0.288 (8)
10	Froma	0.539 (9)	0.509 (9)	0.549 (12)	0.325 (14)	0.571 (10)	0.632 (10)	0.337 (11)

[2] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* , 2(3):27:1–27:27, 2011.

[3] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[4] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[5] V. Jampani, R. Gadde, and P. V. Gehler. Video propagation networks. In *Proceedings of IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, 2017.

[6] A. Khoreva, F. Perazzi, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, 2017.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105. 2012.

[8] T. N. Le, Y. T. Le, M. T. Tran, and A. D. Duong. Essential keypoints to enhance visual object recognition with saliency-based metrics. In *Proceedings of International Conference on Control Automation Robotics Vision (ICARCV)*, pages 111–116, 2014.

[9] T.-N. Le and A. Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In *Proceedings of British Machine Vision Conference (BMVC)*, 2017.

[10] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *IEEE International Conference on Computer Vision, ICCV*, pages 1995–2002, 2011.

[11] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 678–686, June 2016.

[12] T. V. Nguyen and L. Liu. Salient object detection with semantic priors. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[13] T. V. Nguyen and J. Sepulveda. Salient object detection via augmented hypotheses. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2176–2182, 2015.

[14] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 724–732, 2016.

[15] 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.

[16] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[17] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, Aug. 2004.

[18] Y. H. Tsai, M. H. Yang, and M. J. Black. Video segmentation via object flow. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3899–3908, 2016.

[19] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[20] J. Yang, B. Price, S. Cohen, H. Lee, and M. H. Yang. Object contour detection with a fully convolutional encoder-decoder network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 193–202, June 2016.

[21] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

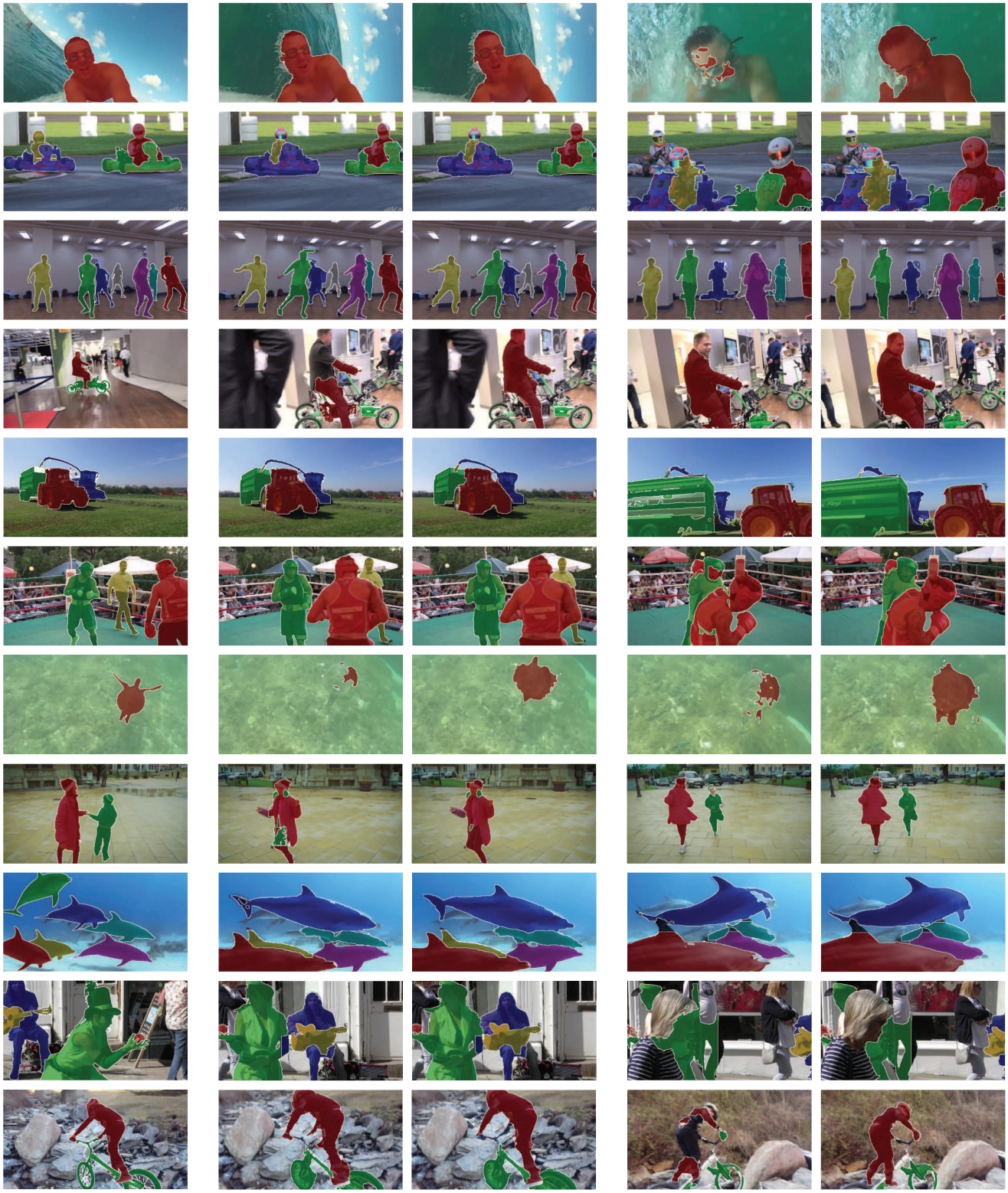

Figure 5. The visualization results of our proposed IRIF method on the DAVIS 2017 dataset. (left to right): the first video frame with the ground truth label, frame t without refinement, frame t with refinement, frame t' without refinement, and frame t' with refinement. The final refinement is done with human re-identification based PSPNet blending, rare instance consideration and boundary snapping. The ground truth of the certain video frame is not publicly available. Our final results significantly track and segment the key objects as annotated in the first frame.