ALBA: Reinforcement Learning for Video Object Segmentation

Shreyank N Gowda*
S.Narayana-Gowda@sms.ed.ac.uk
Panagiotis Eustratiadis*
P.Eustratiadis@sms.ed.ac.uk
Timothy Hospedales
t.hospedales@ed.ac.uk
Laura Sevilla-Lara
l.sevilla@ed.ac.uk

School of Informatics,
University of Edinburgh, UK

Abstract

We consider the challenging problem of zero-shot video object segmentation (VOS). That is, segmenting and tracking multiple moving objects within a video fully automatically, without any manual initialization. We treat this as a grouping problem by exploiting object proposals and making a joint inference about grouping over both space and time. We propose a network architecture for tractably performing proposal selection and joint grouping. Crucially, we then show how to train this network with reinforcement learning so that it learns to perform the optimal non-myopic sequence of grouping decisions to segment the whole video. Unlike standard supervised techniques, this also enables us to directly optimize for the non-differentiable overlap-based metrics used to evaluate VOS. We show that the proposed method, which we call ALBA outperforms the previous state-of-the-art on three benchmarks: DAVIS 2017 [2], FBMS [20] and Youtube-VOS [27].

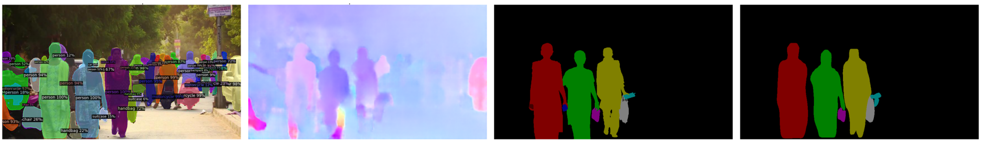

Figure 1: The proposed ALBA for zero-shot video segmentation takes as input the region proposals (left), optical flow (middle left) selects and groups over time the moving ones (middle right), to mimic the ground truth (right).

1 Introduction

Video object segmentation (VOS) aims to segment and track objects within a video. This can be thought of as a grouping problem, where a video is a collection of image regions (pixels, parts or objects) and segmentation is the selection and assignment or grouping of these regions together within and across frames.

Region selection can be informed by appearance cues like, “this object is likely to be moving, because it is a dog"; and temporal features learned from the optical flow could learn things like “this object is moving independently". At the same time, region grouping may also be informed by appearance cues, for example assigning adjacent or similar-looking regions to the same object or leveraging knowledge like person-like regions are more likely to appear *above* bicycle-like regions than vice-versa. Grouping may also benefit from temporal information through alignment of the current frame with the previous frame, as offset by optical flow. Taking such a grouping perspective, we would ideally like to find an assignment of labels to regions that makes a coherent decision about region grouping both spatially within frames and temporally, across frames. However, exhaustively considering the entire joint space of potential assignments to regions is intractable, even for videos just a few seconds long.

In the deep learning era, the majority of studies [8] address the VOS problem at the pixel level, with fully convolutional networks that perform dense pixel-wise classification at every frame. Spatial coherence is maintained through convolution or recurrence [8, 22], and temporal coherence is maintained through recurrence [22], memory [21] or attention [32]. All these methods address the grouping aspect of the VOS problem in an implicit way.A minority of studies have considered explicit grouping based on object proposals, focusing on issues such as: learning good representations to inform the grouping decision (*e.g.* from flow, images, saliency, etc) [11, 31], and efficiently searching the joint space of groupings to find solutions that approximate the global optimum such as via multi-hypothesis tracking [26] or differentiable analogies to the Hungarian algorithm [31].

Despite extensive excellent work in both families of approaches, existing methods have limitations that arise from the near ubiquitous use of supervised learning for training. In particular: (1) For models that sequentially consider individual frames or grouping decisions, training with supervised learning provides a kind of teacher-forcing [1, 15]. That is, from the perspective of any given decision, the preceding decisions benefit from supervision during training in a way that they do not during testing. This creates a mismatch between training and testing conditions termed exposure bias [15]. This results in increased error at testing, since the model has not been trained to make optimal decisions when using its own open-loop predictions as history. (2) The use of supervised learning means that prior work is almost always trained by pixel-wise cross-entropy loss. This provides a strong form of supervision that is easy to differentiate from end-to-end learning. However optimising pixel-wise loss does not necessarily maximise the segment-overlap type of metrics of interest in VOS.

To address these issues, we introduce a reinforcement learning-based method for VOS. Training with reinforcement learning (RL) enables us to (1) Train the model to directly optimize the intersection over union (IoU)-style metrics of interest in VOS, (2) Optimize for a policy that makes the best sequence of *hard* assignment decisions both within and across-frames. Crucially, RL training enables these decisions to be non-myopic, rather than greedy [31]. A non-myopic grouper makes each decision to maximise the expected overall segmentation performance of a video, rather than merely the expected result of an individual grouping decision.

Specifically, we propose a simple and effective framework for segmentation by proposal grouping. Our simple relational [16, 19] architecture takes as input common cues (image, flow) along with the current proposal and the groups so far. It then computes the relation (group or not) between each proposal and available groups. This recurrent assignment of labels to proposals enables efficient joint decision-making about the within-frame grouping – exploiting shape, appearance, and motion cues. Temporally-coherent assignment is further enabled through a very light-weight temporal recurrence.

We call the proposed method ALBA (for *Advanced Learning for Boosted Accuracy*). Without bells and whistles, we observe that RL with ALBA improves 6% over using traditional supervised learning on DAVIS 2017. The final results outperform the state-of-the-art in the DAVIS 2017 [2] validation set, FBMS [20] and the recent large-scale Youtube-VOS [27].

2 Related Work

The video object segmentation problem is studied in different flavours and under different constraints. Depending on the type of labels at training time, we find instance segmentation, semantic segmentation, plenoptic segmentation or motion segmentation. Depending on the input at test time, we find one-shot segmentation (where the initial frame ground truth is given), interactive, or zero-shot. Depending on the type of scenes, we find single-object and multi-object segmentation. While ideas may be applicable to more than one of these problem categories, for simplicity we focus on the most relevant to us, which are those that address the same problem definition as us: zero-shot, multiple-object, instance segmentation.

Zero-Shot Object Segmentation in Video. The zero-shot (also referred to as "unsupervised") multi-object setting in video segmentation provides specific challenges. The first is the selection of the right objects to track. While in one-shot (or "semi-supervised") the relevant objects are given, in the zero-shot they need to be discovered. Research efforts have taken a variety of approaches to address this issue, including the use of annotated eye-tracking data [24] that contains information about the importance of objects, modelling saliency [17] in the appearance, or learning attention [32]. The second is the tracking of multiple instances of objects, where object identities are to be matched over time. While the single-object segmentation problem can be solved with little or no temporal information [13], matching objects across frames requires more sophisticated temporal modelling. Recent efforts have leveraged recurrent networks [17, 22], graphs [23] or attention [28]. While these efforts have yielded great progress in the benchmarks over the years, we may say that most of the focus has been on enriching representations through elaborate models, keeping the optimization process fixed, as supervised learning. Instead, our approach is to use a fairly simple representation, and focus on the optimization.

Reinforcement Learning in Vision and Video. Reinforcement learning (RL) is conventionally used for tackling sequential decision making problems such as robot control or game-playing [14]. In recent years RL has increasingly been exploited in computer vision, where sequential decision problems also arise. For example: deciding which sequence of augmentation operators to apply in data augmentation pipelines [4], sequence of image processing operators to apply in image restoration [29], or which sequence of words to predict in image captioning [12]. Most related to our problem, RL has often been exploited in object tracking [3, 30], where decisions about object locations and identities at time t obviously affects the inference about these objects at time $t+1$. In the VOS problem, a related sequential decision-making problem arises as prior grouping decisions within a frame or at earlier frames affect the labelling decision for the subsequent regions. However, very few studies have attempted to apply RL to VOS. The main example thus far [6] essentially uses RL to solve a single object semi-supervised bounding-box tracking problem and then performs conventional segmentation within the tracked bounding box. In contrast, our framework performs multi-object VOS, directly optimised by RL. Please note that despite the similar name, [5] is completely unrelated as it uses a fixed segmentation module to improve RL-game playing, rather than RL to improve segmentation. Besides enabling learning of sequential decision policies RL also enables the optimization of non-differentiable rewards. This has been exploited in other tasks such as captioning [12] (to optimize language metrics) and tracking [3, 30] (to optimize IoU metrics). In this paper we exploit RL-based training to optimize such overlap-based metrics for video segmentation, rather than conventional pixel-wise cross-entropy.

3 Method

Our overall goal is to process a sequence of video frames I_t with multiple moving objects, and generate a sequence of segmentation tensors M_t that label each pixel with the consistent instance identity. So that each moving object in the video is tracked by a tube in M_t. Our framework generates object proposals and optical flow fields at every frame using off-the-shelf methods. These are fed to our selection network, that rejects stationary proposals and accepts moving proposals. Our assignment network then considers all proposals in one frame given the previous frame's segmentation, and makes a joint decision for how to group the current frame's proposals. Segmenting a video can thus be seen as a sequence of within-frame and across-frame grouping decisions. As a sequential decision problem, we train our model with reinforcement learning to optimise the total prediction-groundtruth overlap for the whole video. We explain each of these components in the following sections.

3.1 Architecture

The architecture is summarized in Fig. 2. We process a sequence of images I_t. For each frame t we generate an associated optical flow image (u_t, v_t) along with a set of object proposals $\{p_i, f_i, b_i\}_t$, each described by a mask p_i, deep image feature f_i and bounding box b_i. As output, we produce a segmentation tensor M_t at every frame that labels each pixel with a persistent object ID or as background. For image size $w \times h$ and max number of objects K_{max} then $M_t \in \{1,0\}^{w \times h \times K_{max}}$. We can then project M_t to a 2D segmentation in $\{0, \dots, K_{max}\}^{w \times h}$.

Selection Network. We run an off-the-shelf category-agnostic proposal generator with low threshold to ensure all objects are considered with low false-positive rate. Our selection network then considers each proposal i in turn, and accepts it as moving, or rejects it as stationary/background. It consists of a convolutional encoder module that inputs the depth concatenated RGB image, flow image, and proposal mask. The flow images are cropped to focus on the current region using its bounding box b_i. After processing by the selection encoder and pooling out spatial information, the appearance feature vector f_i is fused before further processing by a MLP to produce a single binary output. This setup enables the selection network to detect moving foreground objects by performing spatial reasoning about shape (from the proposal), motion (from flow) and appearance (RGB, deep feature), while avoiding distraction (due to cropping).

Assignment Network. The goal of the assignment network is to take the sequence of selected proposals $\mathbf{p}_t = \{p_{i,t}\}$ for each frame t and group them into a consistent set of objects over time. If we denote the grouping label of proposal i as m_i, then the goal at each frame is to make a joint decision about $p(M_t | \mathbf{p}_t, I_t, u_t, v_t, \mathbf{f}_t, M_{t-1})$ where $\mathbf{f}_t$ contains the appearance features for all selected proposals and $M_t = \{m_i\}_t$ denotes the labels of all proposals in the

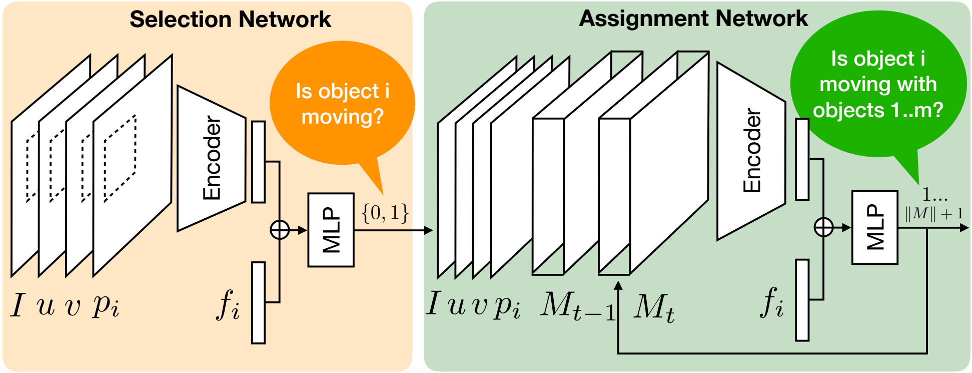

Figure 2: Overview of the architecture. The selection network discriminates between moving and non-moving objects proposals (p_i). Given a moving object i, the assignment network assigns an ID based on coherent movement, appearance and shape, IDs assigned to objects in the previous frame M_{t-1}, and IDs assigned so far in the current frame M_t.

frame, or equivalently its segmentation. To make this joint inference, we use the tractable recurrent factorisation:

$$p(M_t|\mathbf{p}_t, I_t, u_t, v_t, \mathbf{f}_t, M_{t-1}) = \prod_i p(m_{i,t}|M_{<i,t}, p_{i,t}, I_t, u_t, v_t, f_{i,t}, M_{t-1}) \tag{1}$$

where $M_{<i,t}$ indicates the segmentation/grouping decisions in frame t so far, prior to region i. That is, we consider each proposal i in turn and assign it an object label based on its shape and appearance $(p_{i,t}, f_{i,t})$, the shared conditioning information (I_t, u_t, v_t), as well as the labels assigned to each proposal in the frame thus far $M_{<i,t}$ and labels in the prior frame M_{t-1}.

To define the probability $p(m_{i,t} = k|p_{i,t}, \dots)$ that a proposal $p_{i,t}$ is assigned to object $k \in \{1 \dots K_{max}\}$, we take a softmax over logits $l^k_{i,t}$. Our assignment net predicts each logit $l^k_{i,t}$ by a Siamese relational network [16, 19] that 'compares' each proposal i with putative grouping target k. Specifically, the assignment network contains a CNN encoder module that inputs the depth concatenated images $(I_t, u_t, v_t, p_{i,t}, M^k_{t-1}, M^k_{<i,t})$ where $M^k_{<i,t}$ and M^k_{t-1} denote taking the k-th object slice out of the corresponding tensors. These are processed and spatially pooled, before fusing with the appearance feature f_i by concatenation, and then fed into an MLP module that generates the logit $l^k_{i,t}$. After the assignment network generates all the logits for one proposal we have defined $p(m_{i,t}|\dots)$ and we choose the max. Once an assignment decision is made, we update $M_{i,t}$ which is fed to the next iteration of Eq. 1. The proposed masks are sorted by the confidence level of the proposal generator [7]. In this way, we sequentially label each region in each frame, and each frame in the video.

3.2 Training

We pre-train the selection and assignment networks with supervised learning as warm up, and then train the assignment network with reinforcement learning. Fig. 3 shows the architecture of the assignment and selection networks.

Ground Truth Generation. In order to pre-train our network with supervised learning, we need ground truth labels for selection and assignment/grouping of the initial mask proposals. While the segmentation datasets provide ground truth assignments at the pixel level,

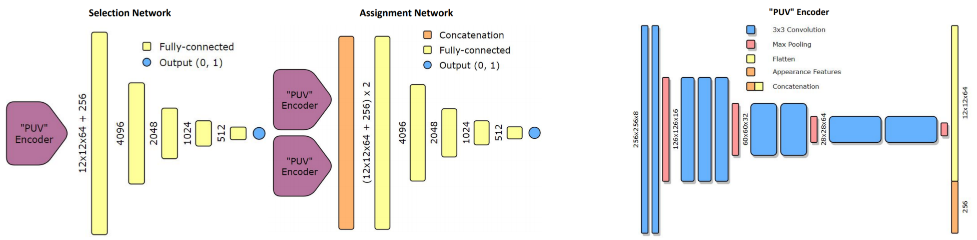

Figure 3: Left: Selection and assignment networks. Right: The "PUV" encoder network used as a module in the selection and assignment networks."PUV" stands for the depth-concatenated mask proposals (P) with the optical flow (UV).

this only helps us partially. For supervised learning at the proposal level, we need the optimal assignment given a set of mask proposals. Since this optimal assignment is costly to find, we use an approximation based on defining a greedy oracle with respect to the ground truth. We generate the selection network ground truth of whether objects are moving or not by iterating over the proposed masks and selecting those that overlap with any object instance in the ground truth more than 0.2. We generate the ground truth for assignments by iterating over the selected masks and matching those that improve the current IoU of any the object instance. Note that this is not a perfect solution as different mask ordering (by confidence in our case) may lead to different solutions, and no ordering may be a global optima. We report the performance of this greedy oracle in the experiments. But note that this only applies to our supervised stage, our RL stage can potentially improve on this.

Supervised Pre-Training. Given the generated "ground-truth", we train the selection net and assignment net by supervised learning with log-loss and cross-entropy respectively. After training, our selection network has 90% accuracy (DAVIS 2017) on selecting true moving objects for subsequent consideration by the assignment network. We use a fixed 20 epoch pre-training.

Training with Reinforcement Learning. RL algorithms are formalized by defining their state space, action space, transition function and reward. From this perspective, our state space consists of the set inputs to the assignment network including common (image, flow, etc) and specific to the current proposal (current mask, appearance, etc). Our action space is the set of possible grouping assignments for the current proposal. The transition function updates the state to provide the next proposal (within frame boundaries), and also next frame (across frame boundaries). It also updates the 'grouping so far' $M_{<i,t}$ channel of the state with the results of the previous grouping action. Finally, we need to define the reward. The reward R^v for video v is the discounted sum of rewards per frame $R^v = \sum_t \gamma^t r_t^v$, where $r_t^v = \text{overlap}(M_t, GT_t)$, and $\gamma = 0.99$. Overlap is quantified by $J\&F$-mean metric. We then use vanilla policy gradient [25] training to optimize for the expected reward across all videos.

Implementation and Training Details. We choose as mask proposal generator the widely used MaskRCNN [7], without fine-tuning it. Note that this network is trained in the COCO dataset, where categories sometimes do not overlap with those found for example on the DAVIS 2017 dataset. Therefore, we use a low-confidence threshold of 0.05, to avoid discarding potential valid object instances. Even if the class is not correct because it was not in the training set, MaskRCNN tends to detect object-like regions. For the optical flow estimation we use the PWC-Net [18]. We enlarge the training set by doing some basic data augmentation, flipping all images horizontally. We train each of the selection and assignment networks for 30 epochs and use Adam optimizer with a learning rate of 1e-4. We use a batch size of 16 and reshape the flow and mask features to 256x256. For the RL part, we use a single agent that maximises the discounted reward. The encoder has 9 convolutional layers and 4 max-pooling layers and results in an output that corresponds to high level features. The MLP in the selection and assignment network consists of 5 FC layers each.

4 Experiments

We show experimental results of the proposed ALBA network. First, we test each of the proposed components in an ablation study, and observe that they all have a significant impact in the performance. In particular, training the network with RL is the component that improves most. We then compare the proposed method to current state-of-the-art methods, and observe that it outperforms all published previous work.

Datasets and Settings. We use three datasets for our experiments. We use the *DAVIS 2017* [2] since it is the most widely used dataset for zero-shot, multi-object, instance segmentation benchmark. It contains 60 training and 30 validation sequences with 4209 and 1999 frames respectively. The performance metrics are region similarity or intersection over union J, and the boundary accuracy F. We use J&F-Mean as the overlap metric for reward on all three datasets. We also use the very recent *YouTube-VOS* [27], which is the the largest video object segmentation dataset to date, and a very promising benchmark. While it does not include an official benchmark for the zero-shot segmentation problem, it can be easily adapted by simply not using the first frame as input. Previous zero-shot methods [22] have also used that approach. The YouTube-VOS includes 3,471 and 474 videos in the training and validation set respectively. Finally, we use the *FBMS* [20] dataset, which contains 59 video sequences, 29 are training and 30 for testing.

Baselines. For the ablation study we use the output from the object proposal generator (MaskRCNN [7]) as a lower bound baseline. As an upper bound we use the generated ground truth described in Sec. 3.1. For the comparison to current state-of-the-art we simply use the top-performing methods from the DAVIS 2017 benchmark [1], which include AGS [24], RVOS [22] and PDB [17]. It is worth noting that AGS is the top-performing of the three, but uses additional annotations of the saliency in DAVIS 2017. None of the other two methods or our own makes use of this additional labels.

4.1 Ablation Study

We test the different components of the proposed ALBA network and show the results in Table 1. We start with the selection network (referred to as S in the table), which discards non-moving objects. The assignment is done by simply assigning a new object identity to each proposed mask. We observe that the selection network improves 4.4% over the vanilla MaskRCNN results. While DAVIS 2017 does not penalize selecting non-moving objects, we attribute this improvement to reducing the number of possible mistakes at the assignment stage. We then show the results of adding a simple assignment network that takes in two proposals and predicts whether they belong to the same object. This assignment network is denoted A in the table. Compared to the naïve assignment, the simple assignment network

[1] https://davischallenge.org/davis2017/soa_compare.html

Method	J&F-Mean	J-Mean	J-Recall	J-Decay	F-Mean	F-Recall	F-Decay
MRCNN	38.9	37.0	42.3	0.3	40.9	43.0	2.5
S	43.3	41.1	46.7	**-0.3**	45.5	47.8	5.0
S+A	49.6	49.3	52.7	4.3	49.9	51.8	2.5
S+A+T	52.9	52.4	55.1	4.1	53.4	54.6	**2.1**
S+A+T+RL	**58.4**	**56.6**	**63.4**	7.7	**60.2**	**63.1**	7.9
Oracle	64.1	62.9	72.0	-2.0	65.3	73.1	1.7

Table 1: **S**: Selection network, **A**: Assignment Network, **T**: Temporal information, **RL**: Optimized with Reinforcement Learning. Results of different components of the proposed network on DAVIS 2017 validation set.

improves significantly. We then add the more sophisticated assignment network from Fig. 2, which includes the temporal component (T) M_{t-1} and observe that this improves results further, due to object labels being more consistent over time. Finally, we train the assignment network using RL, as described in Sec. 3.2, and observe a 5.5% increase in performance. Since the architecture is unchanged in this step, this is attributable to the switch form greedy (supervised) to non-myopic optimisation of assignment as well as optimisation of the target J&F metric. This large jump shows the value of our contribution in terms of introducing RL to the VOS problem.

We also show some sample qualitative results in Fig. 4. It is worth noting that the selection network correctly cleans up the initial proposals, for example in the case of crowded scenes like the breakdancing and the biking scene.

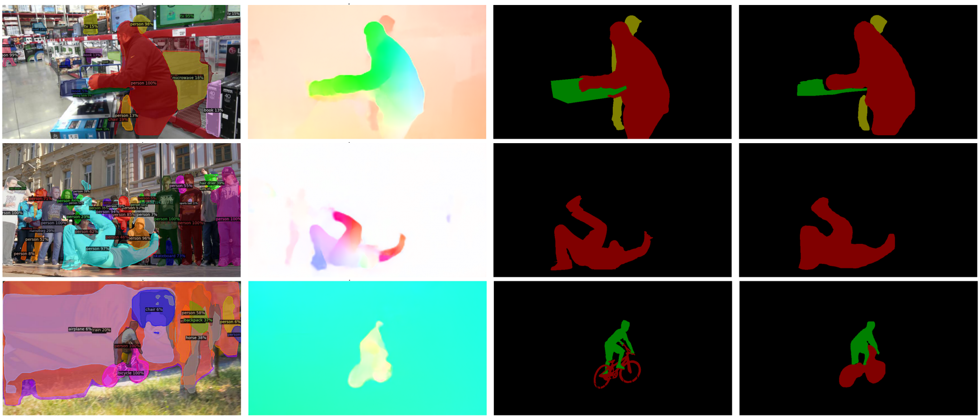

Figure 4: Zero-shot segmentation results. From left to right: original image with MaskR-CNN results super-imposed, optical flow estimation, ground truth, results from our method.

4.2 Comparison to State-of-the-art

We now compare the full proposed ALBA method to the current state-of-the-art . The numerical results are shown in Table 2 for DAVIS2017, Table 3 for FBMS and Table 4 for Youtube-VOS respectively. The results show that ALBA consistently outperforms alternatives across different datasets, although they come from different sources and thus expose

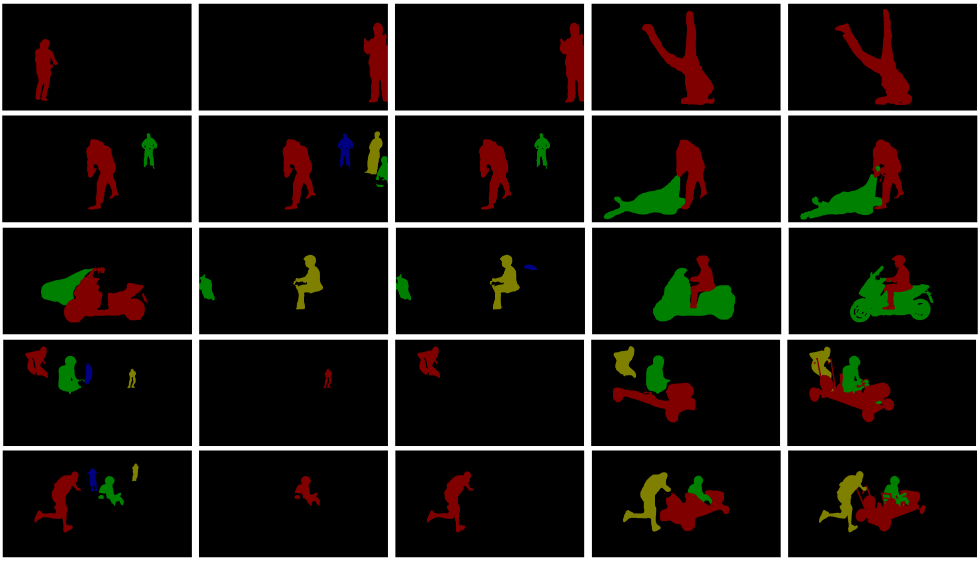

Figure 5: Qualitative comparison to state-of-the-art on DAVIS 2017. From left to right: AGNN, PDB, AGS, ALBA , ground truth.

different challenges. For example, the criteria for different datasets are slightly different in each dataset; objects that move together are grouped together in FBMS, while objects that are semantically different are distinct in the others – creating differences for person plus horse type examples. Meanwhile Youtube-VOS introduces unseen object classes between training and testing. Qualitative results comparing ALBA to state-of-the-art alternatives on the DAVIS dataset are shown in Fig. 5.

Method	Add. Data	Post Process.	J&F Mean	J Mean	J Recall	J Decay	F Mean	F Recall	F Decay
PDB [17]	✗	✓	55.1	53.2	58.9	4.9	57.0	60.2	6.8
RVOS [22]	✗	✗	41.2	36.8	40.2	0.5	45.7	46.4	1.7
ALBA	✗	✗	**58.4**	**56.6**	**63.4**	7.7	**60.2**	**63.1**	7.9
AGS [24]	✓	✓	57.5	55.5	61.6	7.0	59.5	62.8	9.0

Table 2: Comparison to other state-of-the-art on the DAVIS 2017 validation set.

Method	MAT [32]	MBN [10]	PDB [17]	IET [9]	ALBA
J-mean	76.1	73.9	74.0	71.9	**77.6**
F-score	-	83.2	**84.9**	82.8	84.4

Table 3: Results on FBMS dataset.

Method	J-seen	J-unseen	F-seen	F-unseen
RVOS [22]	44.7	21.2	45.0	23.9
ALBA	**53.8**	**34.4**	**51.6**	**35.8**

Table 4: Results on Youtube-VOS dataset.

5 Conclusion

We considered the zero-shot video object segmentation task as a selection and grouping problem over generic object proposals, and discussed how training such a model with RL enables better global decision-making and direct optimization of overlap metrics. We showed that our grouping architecture surpasses previous approaches when trained with RL. We believe that this is the first demonstration of RL in VOS and hope that it leads others to leverage this technique for VOS, VIS and related tasks in future.

References

[1] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, 2015.

[2] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019.

[3] Boyu Chen, Dong Wang, Peixia Li, Shuang Wang, and Huchuan Lu. Real-time 'actor-critic' tracking. In *ECCV*, 2018.

[4] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *CVPR*, 2019.

[5] Vikash Goel, Jameson Weng, and Pascal Poupart. Unsupervised video object segmentation for deep reinforcement learning. In *NeurIPS*, 2018.

[6] Junwei Han, Le Yang, Dingwen Zhang, Xiaojun Chang, and Xiaodan Liang. Reinforcement cutting-agent learning for video object segmentation. In *CVPR*, 2018.

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

[8] Suyog Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. In *CVPR*, 2017.

[9] Siyang Li, Bryan Seybold, Alexey Vorobyov, Alireza Fathi, Qin Huang, and C-C Jay Kuo. Instance embedding transfer to unsupervised video object segmentation. In *CVPR*, 2018.

[10] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *ECCV*, 2018.

[11] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *ECCV*, 2018.

[12] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, 2017.

[13] Kevis-Kokitsi Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 02 2015.

[15] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016.

[16] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Timothy P. Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, 2017.

[17] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, 2018.

[18] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.

[19] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.

[20] T.Brox and J.Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.

[21] P. Tokmakov, K. Alahari, and C. Schmid. Learning video object segmentation with visual memory. In *ICCV*, 2017.

[22] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. RVOS: End-to-end recurrent network for video object segmentation. In *CVPR*, 2019.

[23] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, 2019.

[24] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *CVPR*, 2019.

[25] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 1992.

[26] Shuangjie Xu, Daizong Liu, Linchao Bao, Wei Liu, and Pan Zhou. MHP-VOS: Multiple hypotheses propagation for video object segmentation. In *CVPR*, 2019.

[27] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019.

[28] Zhao Yang, Qiang Wang, Luca Bertinetto, Song Bai, Weiming Hu, and Philip H.S. Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, 2019.

[29] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *CVPR*, 2018.

[30] Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi. Action-decision networks for visual tracking with deep reinforcement learning. In *CVPR*, 2017.

[31] Xiaohui Zeng, Renjie Liao, Li Gu, Yuwen Xiong, Sanja Fidler, and Raquel Urtasun. DMM-Net: Differentiable mask-matching network for video object segmentation. In *CVPR*, 2019.

[32] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*, 2020.