

# TRICKVOS: A BAG OF TRICKS FOR VIDEO OBJECT SEGMENTATION

*Evangelos Skartados<sup>1</sup>\*, Konstantinos Georgiadis<sup>1</sup>\*, M. Kerim Yucel<sup>2</sup>\**

*Koskinas Ioannis<sup>1</sup>, Armando Domi<sup>1</sup>, Anastasios Drosou<sup>1</sup>, Bruno Manganelli<sup>2</sup>, Albert Saà-Garriga<sup>2</sup>*

<sup>1</sup> CERTH, Information Technologies Institute, Thessaloniki, Greece

<sup>2</sup> Samsung Research UK

## ABSTRACT

Space-time memory (STM) network methods have been dominant in semi-supervised video object segmentation (SVOS) due to their remarkable performance. In this work, we identify three key aspects where we can improve such methods; i) supervisory signal, ii) pretraining and iii) spatial awareness. We then propose TrickVOS; a generic, method-agnostic bag of tricks addressing each aspect with i) a structure-aware hybrid loss, ii) a simple decoder pretraining regime and iii) a cheap tracker that imposes spatial constraints in model predictions. Finally, we propose a lightweight network and show that when trained with TrickVOS, it achieves competitive results to state-of-the-art methods on DAVIS and YouTube benchmarks, while being one of the first STM-based SVOS methods that can run in real-time on a mobile device.

**Index Terms**— Video Object Segmentation, Pretraining, Space-time Memory Networks

## 1. INTRODUCTION

Semi-supervised video object segmentation segments objects, where object mask annotations are given only for the first frame. SVOS is the backbone of many applications, such as surveillance, autonomous driving and video manipulation. It is a difficult task, where robustness against occlusions, shape changes, motion and distractors are necessary, and performing well irrespective of the object classes is imperative. In recent years, space-time memory networks have gained traction [1, 2, 3, 4], as they have made significant advances. Nevertheless, there is still a considerable room for improvement.

In this paper, we aim to improve STM networks by introducing a generic, method-agnostic bag-of-tricks. We first identify three points relatively unexplored in the literature; i) supervisory signals, ii) pretraining and iii) spatial constraints. We then propose TrickVOS that includes solutions for each of these points. The contributions of TrickVOS are as follows; i) a training loss that leverages the structural information available in the ground-truth masks, with an emphasis on the tracked object regions, ii) a simple pretraining regime that addresses the problem of not being able to effectively pretrain decoders and iii) a cheap tracker that filters the model predictions based on historical motion patterns. Each trick

improves the results, and their combination leads to further improvements. Finally, we demonstrate that TrickVOS makes a lightweight network competitive to significantly larger/expensive methods on multiple SVOS benchmarks. TrickVOS is visualized in Figure 1.

## 2. RELATED WORK

**Semi-supervised Video Object Segmentation.** SVOS methods span a large spectrum, where per-mask online finetuning [5], flow-based [6], detection-aided [7] and matching-based methods are prominent examples. STM based methods have been the dominant approach, where a feature correspondence problem is solved between the current frame and the frames stored in the memory bank to predict the object masks [1, 2, 3, 4]. Follow-up works have focused on increasing efficiency [4, 2], accuracy on long-videos [8], bounding the memory limit [3, 9] and eliminating spatial redundancy [10, 11]. Despite these advances, there are unexplored areas where STM based SVOS methods can be further improved.

The tricks of TrickVOS can be easily plugged into other methods. To the best of our knowledge, we are the first to propose an enhanced loss function and a decoder pretraining regime to improve the accuracy of SVOS methods. Similar ideas for spatial-filtering have been explored in [11, 10], but ours have key differences; we i) do not use a flow network as [11] or ii) do not perform filtering on the feature maps as [11, 10]. TrickVOS paves the way for accurate STM-based SVOS models that can run in real-time on mobile devices.

## 3. TRICKVOS

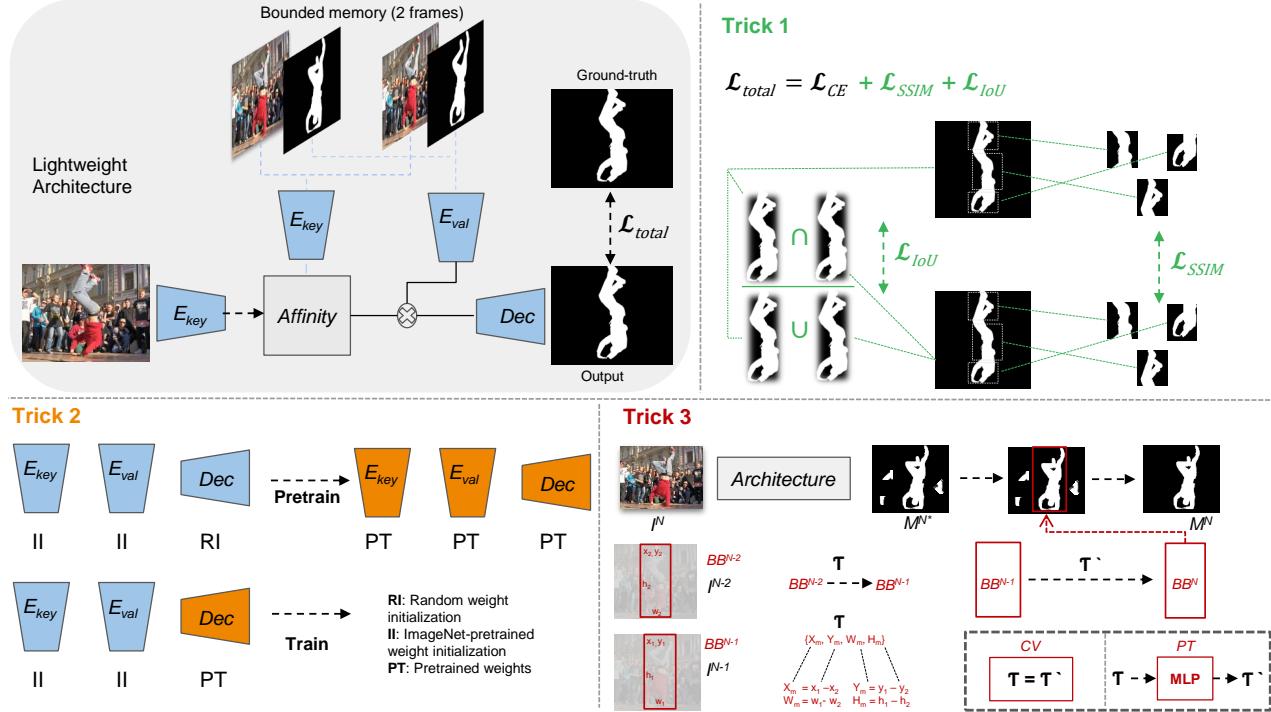
### 3.1. SVOS Preliminaries

Given a set of  $K$  object masks  $M_{obj} \in \mathbb{N}^{HW}$ ,  $M = \{M_1, \dots, M_k\}$  in the first frame  $I^1 \in \mathbb{R}^{HW \times 3}$  of an N-frame long sequence  $I = \{I^1, \dots, I^N\}$ , SVOS aims to track these objects in the sequence. We now identify key issues and propose solutions to improve the performance of SVOS methods.

### 3.2. Trick 1: A Better Loss Function

**Problem.** SVOS is a dense classification task, where each pixel is classified as background or an object. Therefore, most methods train with a form of cross-entropy (CE) loss [2]. However, SVOS is highly sensitive to object shapes, which requires structural information. Structural information also provides boundary awareness, where most SVOS prediction errors tend to occur. Additionally, during training, foreground

\* The first three authors contributed equally.



**Fig. 1.** TrickVOS is formed of three tricks that improve semi-supervised video object segmentation; Trick 1 introduces a loss that promotes foreground focus and structure-awareness, Trick 2 presents a simple decoder pretraining scheme and Trick 3 filters predictions via a cheap tracker with two variants, constant velocity (CV) and parametric tracker (PT).

focus is highly desirable for accuracy and rapid convergence. Neither structural information, nor foreground-focus are provided by CE since it considers all pixels as independent.

**Our solution.** Our first trick aims to provide structural information and foreground-focus during training. For the former, a natural choice is the SSIM metric [12]. SSIM can be used as a loss function to promote structural similarity between predictions and ground-truths, and since its patch-level nature provides rich context information, it might help the network to model the appearance/shape changes more effectively. For the latter, we hypothesize that the Jaccard-index, namely Intersection-over-Union (IoU), can be helpful as it measures the alignment between the predicted and the ground-truth masks. The alignment of the masks naturally focus on the foreground, therefore it meets our criteria. Various differentiable IoU formulations are used in many tasks [13], including SVOS [14], as loss functions. Using the IoU loss of [13], we define the total loss as [15]

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{IoU} \quad (1)$$

where \$\lambda\_1 = \lambda\_2 = \lambda\_3 = 1\$. This formulation is shown to be effective in salient object detection [15], which is somewhat similar to SVOS. See Figure 1 for a diagram of Trick 1.

### 3.3. Trick 2: Decoder Pretraining

**Problem.** Using ImageNet-pretrained encoders in dense prediction tasks is a ubiquitous practice, as it often improves the downstream accuracy. A similar pretraining for decoders is not straightforward; decoders directly produce the final result, which differs for various tasks (i.e. depth vs saliency).

Therefore, a decoder trained on a task might not be a good initialization for another task. Self-supervised learning [16] and generative pretraining [17] are shown to beat random initialization for decoders, but none of them target SVOS.

**Our solution.** Our second trick is probably the simplest one; we simply re-use a decoder that is already trained on our task. We train a model through all the stages and save the decoder weights. Later, we train a model with random weight initialization, except we initialize the decoder weights from the previous training. "Why not train the whole model for longer instead?", one might ask. Alternatively, one might question as to why we do not do the pretraining for the whole model, not just the decoder. We observe that in both cases, the model overfits to DAVIS and performs poorly on YouTube. Our decoder pretraining, however, consistently improves on all datasets (see Table 1). We test alternative strategies [16] but observe poor results. See Figure 1 for a diagram of Trick 2.

### 3.4. Trick 3: Prediction Filtering

**Problem.** STM-based methods perform global feature matching between current and memory frames, which might lead to erroneous matches, especially when objects of same types/appearance are present in the frame. Some alternatives tackle this issue up to a degree, but they require relevant architectural changes since they predict bounding boxes on features [10] or use expensive optical flow [11].

**Our solution.** Our final trick is a tracker that filters the predictions. The core idea is simple; we track a bounding box throughout the video that encapsulates an object mask and fil-

**Algorithm 1** Trick 3 pseudocode.

```

1: function TRICK3( $I, M$ )           > Inputs: Frames I and predicted masks M.
2:    $C = 1$ 
3:   while  $C \neq \text{len}(I)$  do
4:     if  $C < 3$  then
5:        $BB^C = \text{init}(M^C)$           > Initialize bbox from mask.
6:        $C += 1$ 
7:       continue
8:     else
9:        $\mathcal{T} = BB^{C-1} - BB^{C-2}$       > Historical motion pattern.
10:      if CV then
11:         $\mathcal{T}' = \mathcal{T}$               > Constant Velocity variant.
12:      else
13:         $\mathcal{T}' = \text{MLP}(\mathcal{T})$       > Parametric Tracker variant.
14:         $BB^C = \mathcal{T}' + BB^{C-1}$       > Apply motion to predict the bounding box.
15:         $AM^C \leftarrow BB^C$             > Attention map from bounding box.
16:         $M'^C = AM^C \times M^C$        > Filtered mask.
17:         $BB^C = \text{init}(M'^C)$        > Reinitialize bbox from mask.
18:         $C += 1$ 
19:       $M' = \{M'^1, \dots, M'^C\}$ 
20:    return  $M'$                   > Output: Filtered masks.

```

ter out predictions outside the area of the bounding box. We propose two variants of our tracker, Constant Velocity (CV) and Parametric Tracker (PT). Trick 3 is shown in Figure 1.

**Variant 1.** CV is extremely simple; it is not a tracker at all. We simply assume constant velocity for objects; the motion  $\mathcal{T}$  between the previous two bounding boxes is used to calculate the bounding box of the next frame. This bounding box is converted to an attention map, multiplied by the predicted mask to generate the filtered mask. Note that  $\mathcal{T}$  is defined as the change in top left coordinates, width and height.

**Variant 2.** PT extends CV by using an MLP to predict the next bounding box. Instead of assuming a constant  $\mathcal{T}$ , PT feeds the motion between two previous frames into a 5-layer MLP (with GeLU activations in between) to predict the next bounding box. Both CV and PT offer a great deal of flexibility, as CV/PT can be plugged into any method or architecture. See Algorithm 1 for a detailed breakdown of Trick 3.

## 4. EXPERIMENTS

### 4.1. Experimental Details

**Architecture.** We adopt the STCN architecture [2] with key changes; we i) fix the memory size to two frames (first and last) to avoid memory explosion, ii) replace key/value encoders with MobileNetV2 [18], iii) reduce key dimension to 32, iv) remove ASPP and v) use mask as the only input to the value encoder [9]. These changes make our architecture, denoted as the *baseline* throughout the paper, less accurate than STCN, but improves its runtime performance (see Figure 1).

**Datasets.** DAVIS 2016 [19] is a single-object SVOS benchmark with 30 training and 20 validation videos, where DAVIS 2017 [20] extends it to multi-object with 60 training videos, 30 test and 30 validation videos. YouTube-VOS 2019 [21] is a large-scale multi-object SVOS benchmark with 3471 training videos formed of 65 categories and 507 videos for validation. The validation set includes 26 unseen object categories, allowing model assessment on unseen object categories.

**Metrics.** We use the metrics region similarity  $\mathcal{J}$  and contour accuracy  $\mathcal{F}$ , and their average  $\mathcal{J}\&\mathcal{F}$ . On YouTube, we also report seen and unseen class  $\mathcal{J}\&\mathcal{F}$ . We evaluate our methods on the validation sets of DAVIS and YouTube datasets.

Method	DAVIS'16	DAVIS'17	YouTube-VOS'19
Baseline	88.18	80.86	77.18
+ SSIM	87.81 $\downarrow 0.37$	81.25 $\uparrow 0.39$	78.42 $\uparrow 1.24$
+ IoU	89.04 $\uparrow 0.86$	81.88 $\uparrow 1.02$	79.39 $\uparrow 2.21$
+ Trick 1	88.80 $\uparrow 0.62$	81.97 $\uparrow 1.11$	79.42 $\uparrow 2.24$
+ Trick 2	88.47 $\uparrow 0.25$	81.06 $\uparrow 0.20$	78.07 $\uparrow 0.89$
+ Trick 3 (CV)	88.30 $\uparrow 0.12$	81.17 $\uparrow 0.31$	78.04 $\uparrow 0.86$
+ Trick 3 (PT)	88.31 $\uparrow 0.13$	81.23 $\uparrow 0.37$	78.15 $\uparrow 0.97$
+ Trick 1+2	89.19 $\uparrow 0.99$	82.54 $\uparrow 1.68$	79.94 $\uparrow 2.76$
+ TrickVOS (CV)	89.28 $\uparrow 1.10$	82.56 $\uparrow 1.70$	80.38 $\uparrow 3.20$
+ TrickVOS (PT)	89.29 $\uparrow 1.11$	82.67 $\uparrow 1.81$	80.47 $\uparrow 3.29$

**Table 1.** Ablation study  $\mathcal{J}\&\mathcal{F}$  values. CV and PT indicate constant velocity and parametric variants of our tracker (Trick 3). TrickVOS combines all tricks.  $\uparrow$  and  $\downarrow$  show  $\mathcal{J}\&\mathcal{F}$  increase and decrease compared to the baseline, respectively.

**Training Details.** We follow the training regime of [2], where we first pretrain on static images with synthetic deformations, and then train on YouTube and DAVIS 2017 jointly for 300K iterations with a batch size of 16. We use Adam optimizer with a learning rate of  $1e-5$  and a weight decay of  $1e-7$ . The only difference is that we use a sequence length of 5 instead of 3. Trick 3 PT variant is trained separately, using bounding boxes extracted from the ground-truth masks. In PT training, we further penalize the predicted bounding boxes that are smaller than the ground-truth by increasing their loss weights. We use RTX 3090 GPUs for all experiments.

### 4.2. Ablation Studies

We conduct detailed ablation studies to show the effect of each TrickVOS component. The results are shown in Table 1. We use a single model checkpoint to evaluate on all datasets.

**Trick 1.** We add SSIM and IoU losses over the cross-entropy baseline, and then combine all of them. Table 1 third row shows that IoU increases the accuracy significantly. The second row shows that SSIM improves DAVIS'17 but hurts DAVIS'16. Note that we focus on DAVIS'17 as it is a larger dataset with multiple object annotations, which is more reminiscent of an actual SVOS use case. When we combine all (Table 1 4th row), we see solid improvements; 0.62, 1.11 and 2.24  $\mathcal{J}\&\mathcal{F}$  on DAVIS'16, 17 and YouTube, respectively.

**Trick 2.** Fifth row of Table 1 (Trick 2) shows that we get consistent improvements on all datasets. Note that this result shows that we do not overfit, especially to DAVIS since it is quite smaller compared to YouTube. Improvements on YouTube especially emphasize the benefit of Trick 2.

**Trick 3.** We observe improvements even with the constant velocity variant (Table 1 row 6, Trick 3 CV), which has little computational overhead. With the PT variant, we observe slightly better results compared to CV, but at the cost of a slightly increased performance overhead. Note that we do not explicitly optimize the tracker, therefore the performance overhead can be reduced with optimized implementations.

**All tricks.** An important question remains; can we simply use all tricks at the same time and get further improvement? The results show that Trick 1 and 2 combined perform better on all datasets, compared to using them separately. When all tricks

Method	CC	DAVIS'16			DAVIS'17			YouTube'19					FPS	
		$\mathcal{J} \& \mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J} \& \mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J} \& \mathcal{F}$	$\mathcal{J}_{seen}$	$\mathcal{F}_{seen}$	$\mathcal{J}_{unseen}$	$\mathcal{F}_{unseen}$	DAVIS'16	DAVIS'17
MiVOS‡ [22]	✗	91.0	89.7	92.4	84.5	81.7	87.4	82.4	80.6	84.7	78.2	85.9	16.9	11.2
STM [1]	✗	89.3	88.7	89.9	81.8	79.2	84.3	79.2	79.6	83.6	73.0	80.6	6.30	10.2
STCN‡ [2]	✗	91.7	90.4	93.0	85.3	82.0	88.6	84.2	82.6	87.0	79.4	87.7	26.9	20.2
XMem [8]	✗	91.5	90.4	92.7	86.2	82.9	89.5	85.5	84.3	88.6	80.3	88.6	29.6	22.6
GCNet [23]	✓	86.6	87.6	85.7	71.4	69.3	73.5	73.2	72.6	75.6	68.9	75.7	25.0	<25.0
SwiftNet [4]	✓	90.4	90.5	90.3	81.1	78.3	83.9	77.8	77.8	81.8	72.3	79.5	25.0	<25.0
CFBI [24]	✓	89.9	88.7	91.1	81.9	79.3	84.5	81.0	80.6	85.1	75.2	83.0	5.90	—
RDE-VOS [3]	✓	91.1	89.7	92.5	84.2	80.8	87.5	81.9	81.1	85.5	76.2	84.8	35.0	27.0
Baseline	✓	88.2	87.5	88.8	80.9	77.5	84.2	77.2	76.9	81.0	71.3	79.4	102.1	99.5
Baseline + TrickVOS (CV)	✓	89.3	88.7	89.8	82.6	79.3	85.9	80.3	80.0	83.9	74.5	82.9	91.4	79.3
Baseline + TrickVOS (PT)	✓	89.3	88.7	89.9	82.7	79.4	86.0	80.5	79.5	83.3	75.2	84.0	86.4	76.4
STCN † [2]	✗	91.2	90.1	92.3	84.0	80.6	87.4	80.7	80.2	84.5	75.6	82.7	62.4	48.3
STCN † [2] + Trick 1+2	✗	91.7	90.4	92.9	85.4	82.1	88.8	81.6	81.2	85.5	75.8	83.9	62.4	48.3
STCN † [2] + TrickVOS (PT)	✗	91.8	90.5	93.1	86.1	82.6	89.6	82.8	82.1	86.4	77.2	85.5	45.4	35.1

**Table 2.** Results on the DAVIS 2016/2017 and YouTube-VOS 2019 validation sets. CC indicate constant cost. ‡ indicates models trained with additional data (BL30K). † indicates STCN re-training by us (without ASPP module or BL30K pretraining). are combined, we get the largest improvement on all datasets.

### 4.3. Results

We compare TrickVOS against the state-of-the-art methods. We indicate methods with bounded memory during inference with Constant Cost (CC) [3]. Unlike others [3], we use a single model checkpoint for all dataset evaluations. The results are shown in Table 2. Note that methods with unbounded memory are not practical, as they eventually run out of memory and their FPS decrease. We provide their results, but we mainly compare ours against other constant cost methods.

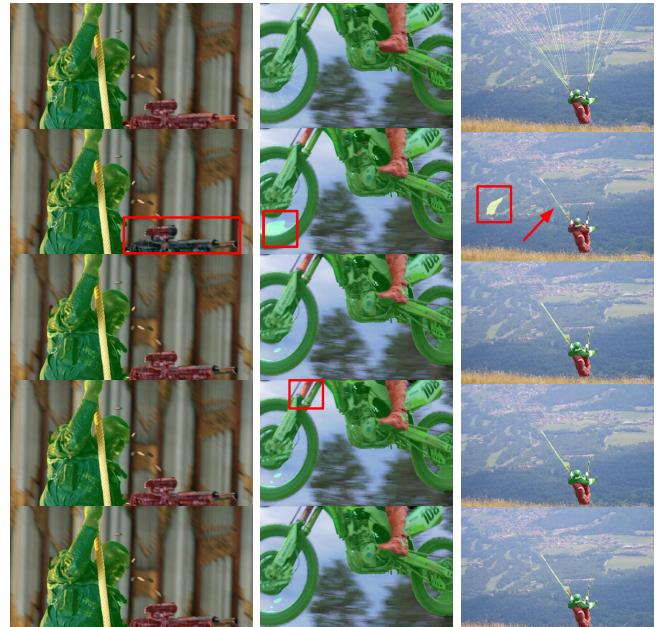
**DAVIS 2016.** Our model performs quite competitively; it is slightly behind RDE-VOS, but it is  $\times 2.5$  faster and has  $\times 32$  fewer parameters (our 1.9M vs their 64M). Our methods comfortably perform in real-time, with minimal reductions (10% and 15% for CV and PT) in runtime compared to the baseline.

**DAVIS 2017.** In multi-object scenarios, the results look even better. We are the second best after RDE-VOS in constant cost methods. Also note that in multi-object scenarios, most of the methods perform significantly slower, but we are affected less than the others; we are now  $\times 3$  faster than RDE-VOS. Our methods are still comfortably in real-time regime, with acceptable reductions in runtime compared to the baseline.

**You Tube-VOS.** We observe the most significant improvements over our baseline on YouTube (3.3  $\mathcal{J} \& \mathcal{F}$ ), which emphasizes the benefit of our tricks as YouTube is the largest and the most diverse of all datasets we experiment with.

**On-device deployment.** TrickVOS unlocks a practical application; mobile deployment. To test on-device performance, we convert our models to tflite and benchmark them using the tflite benchmarking tool. Our method reaches 30+ FPS on a Samsung Galaxy S22 GPU; to the best of our knowledge, this is the first STM-based SVOS method that achieves real-time on a mobile device. Note that we are already faster than others on high-end GPUs (see Table 2); assuming a naive linear scaling of speed from desktop to mobile GPUs and based on our 30 FPS, no other method is likely to perform in real-time.

**Qualitative results.** Figure 2 shows example images from DAVIS'17 validation set; the baseline fails to either track objects (gun in leftmost image) or mispredicts masks (erroneous



**Fig. 2.** Rows show (top→bottom) the ground-truth, baseline, baseline + trick 1, baseline + trick 2 and baseline + TrickVOS predictions. Note the red boxes/arrows showing the erroneous regions, which are gradually improved with our tricks.

predictions in the wheel and on the mountain). Our tricks, when applied, visibly improve such error cases.

**TrickVOS + other methods.** We apply TrickVOS to the original STCN (with no ASPP module or BL30K pretraining) to highlight its plug-and-play nature. For a fair comparison, we re-train it (using authors’ code) with and without TrickVOS; the last three rows of Table 2 show TrickVOS consistently improves the results on all datasets.

## 5. CONCLUSION

We present TrickVOS, a model-agnostic bag of tricks that improves the accuracy of semi-supervised video object segmentation. Each trick introduces improvements, and when combined, they perform even better. TrickVOS, when applied to a lightweight architecture, performs competitively to the state-of-the-art, while running real-time on a mobile phone.

## 6. REFERENCES

- [1] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim, “Video object segmentation using space-time memory networks,” in *International Conference on Computer Vision (ICCV)*, 2019, pp. 9226–9235.
- [2] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang, “Rethinking space-time networks with improved memory coverage for efficient video object segmentation,” *NeurIPS*, vol. 34, pp. 11781–11794, 2021.
- [3] M. Li, Li Hu, Z. Xiong, Bang Zhang, Pan Pan, and Dong Liu, “Recurrent dynamic embedding for video object segmentation,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1332–1341.
- [4] Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai, “Swiftnet: Real-time video object segmentation,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1296–1305.
- [5] Xiaoxiao Li and Chen Change Loy, “Video object segmentation with joint re-identification and attention-aware mask propagation,” in *European conference on computer vision (ECCV)*, 2018, pp. 90–105.
- [6] YH Tsai, MH Yang, and MJ Black, “Video segmentation via object flow,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3899–3908.
- [7] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe, “Premvos: Proposal-generation, refinement and merging for video object segmentation,” in *Asian Conference on Computer Vision (ACCV)*, 2019, pp. 565–580.
- [8] Ho Kei Cheng and Alexander G Schwing, “Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model,” in *European conference on computer vision (ECCV)*, 2022, pp. 640–658.
- [9] Roy Miles, Mehmet Kerim Yucel, Bruno Manganelli, and Albert Saà-Garriga, “Mobilevos: Real-time video object segmentation contrastive learning meets knowledge distillation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10480–10490.
- [10] Bo Miao, M. Bennamoun, Yongsheng Gao, and Ajmal Mian, “Region aware video object segmentation with deep motion modeling,” *arXiv:2207.10258*, 2022.
- [11] H. Xie, H. Yao, S. Zhou, S. Zhang, and W. Sun, “Efficient regional memory network for video object segmentation,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1286–1295.
- [12] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, “Multiscale structural similarity for image quality assessment,” in *ACSSC. Ieee*, 2003, vol. 2, pp. 1398–1402.
- [13] Gellért Mátyus, Wenjie Luo, and Raquel Urtasun, “Deeproadmapper: Extracting road topology from aerial images,” in *International Conference on Computer Vision (ICCV)*, 2017, pp. 3438–3446.
- [14] Ye Yu, J. Yuan, G. Mittal, Li Fuxin, and Mei Chen, “Batman: Bilateral attention transformer in motion-appearance neighboring space for video object segmentation,” in *European conference on computer vision (ECCV)*, 2022, pp. 612–629.
- [15] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jaggersand, “Basnet: Boundary-aware salient object detection,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7479–7489.
- [16] E. A. Brempong, Simon K., Ting C., Niki P., Matthias M., and M. N., “Denoising pretraining for semantic segmentation,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4175–4186.
- [17] Barret Z., Golnaz G., Tsung-Yi Lin, Yin Cui, H. Liu, Ekin DC, and Quoc Le, “Rethinking pre-training and self-training,” *NeurIPS*, vol. 33, pp. 3833–3845, 2020.
- [18] Mark S., Andrew H., M. Zhu, Andrey Z., and LC Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.
- [19] Federico P., Jordi PT, Brian M., Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung, “A benchmark dataset and evaluation methodology for video object segmentation,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 724–732.
- [20] Jordi PT, Federico P., Sergi C., Pablo Arbeláez, Alex SH, and Luc Van Gool, “The 2017 davis challenge on video object segmentation,” *arXiv:1704.00675*, 2017.
- [21] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang, “Youtube-vos: A large-scale video object segmentation benchmark,” *arXiv:1809.03327*, 2018.
- [22] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang, “Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5559–5568.
- [23] Yu Li, Z. S., and Y. S., “Fast video object segmentation using the global context module,” in *European conference on computer vision (ECCV)*, 2020, pp. 735–750.
- [24] Zongxin Yang, Yunchao Wei, and Yi Yang, “Collaborative video object segmentation by foreground-background integration,” in *European conference on computer vision (ECCV)*, 2020, pp. 332–348.