Shifting More Attention to Video Salient Object Detection
Supplemental Material

Anonymous CVPR submission

Paper ID 1853

Abstract

In this document, we include additional materials related to the benchmark, dataset, model, evaluation metrics, and results.

- ***Benchmark.*** *We systematically assess 17 representative VSOD algorithms [3, 4, 7, 9, 11–13, 17–19, 21–25, 27, 29] over seven existing VSOD datasets and the proposed DAVSOD with totally 40K frames, making it the largest-scale benchmark.*
- ***Dataset.*** *We show a complete statistics of existing datasets and provide some representative sample frames of the proposed DAVSOD dataset.*
- ***Model.*** *We describe our implementation details and the training protocol in § 3.*
- ***Metrics.*** *In § 4.1, we describe the details of three metrics,* e.g.*, MAE ($\mathcal{M}$) [15], F-measure ($\mathcal{F}$) [1], and S-measure ($\mathcal{S}$) [5] and the details of metric statistics in Table 4 (*manuscript*).*
- ***Results.*** *We provide more experiment results on existing 7 datasets [7, 8, 10, 12, 14, 16, 23] and the proposed dataset in § 4.3. We refer the reader to the accompanying video attachment (VideoSaliency.mp4 [76MB]) for more details. Due to both limit space and maximum (100MB) file size, the overall dataset will be publicly available in our website.*

All of the model results (17 state-of-the-art and the proposed SSAV model) on 8 datasets (7 existing datasets and the proposed DAVSOD dataset) will be released in our website to boost the development of video salient objection.

1. Benchmark

As shown in Fig. 1, it has witnessed the dramatic development of VSOD modeling, while the community long-term suffered from the lack of a standard representative benchmark. To the best of our knowledge, this work is first one largest-scale VSOD benchmark.

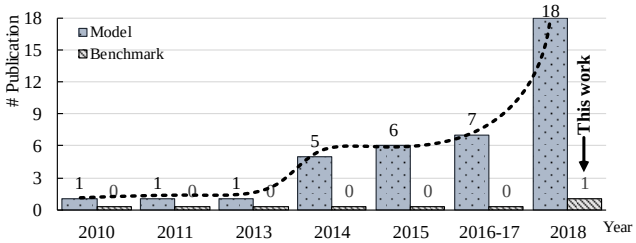


Figure 1: **Number of VSOD papers** on top conferences (*i.e.*, CVPR, ICCV, ECCV) and IEEE Trans. journals over the past 9 years.

2. Dataset

A complete statistics of previous VSOD dataset and the proposed *DAVSOD* dataset in Section 3 (*manuscript*) can be found in Table 1.

In our dataset, the salient objects may be changed at different time (see Fig. 2), which is more realistic and requires a complete video content understanding.

We ask one annotator to provide a brief single-sentence description ($\leq$15 words) for summarizing the main content of each video, after watching the whole video sequence. Meanwhile, the corresponding video and object labels are offered for reference. In Fig. 3 and Table 2 show some examples of the textual description.

As described in Section 3.2 (*manuscript*), the annotators were pre-trained with ten video examples. We also conducted an explicit verification step on each segmented instance to guarantee good quality. Some high-quality annotated examples can be found in Fig. 4. In Fig. 5 and Fig. 6, we visualize a representative frame of each sequence with instance level ground-truth masks and fixation map overlayed in different color. Examples of cases that either passed or were rejected are shown in Fig. 7.

3. Model

3.1. Implementation Details

The base CNN network of PDC model is borrowed from the conv blocks from ResNet-50 [6] and the conv strides of the last two blocks are changed to 1. All the input frame

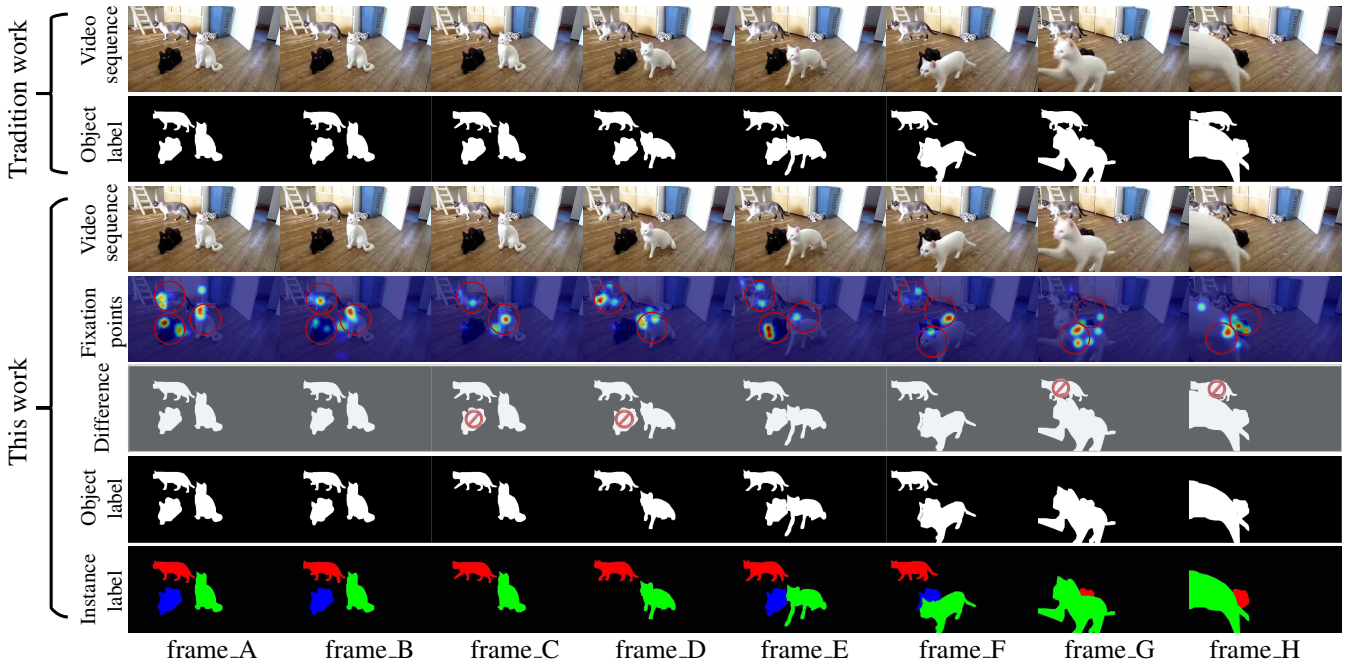


Figure 2: **Example sequence of saliency shift considered in the proposed *DAVSOD* dataset.** Different (5^{th} row) from tradition work which only labels all of the salient object (2^{th} row) via *static* frames, without a *dynamic* human eye-fixation guided annotation methodology. The proposed *DAVSOD* dataset were strictly annotated according to real human fixation record (3^{rd} row), thus reveal real human attention mechanism during dynamic viewing. Zoom-in for details.

images are resized into 473×473 spatial resolution, and $\mathbf{Q} \in \mathbb{R}^{60\times60\times2048}$. Following [18], we set $K = 4$, $C = 512$ and $d_k = 2^k$ ($k \in \{1, \cdots, 4\}$). For the convLSTM in Eq. 2 (*manuscript*), we use a $3 \times 3 \times 32$ conv kernel. The convLSTM^A in Eq. 3 (*manuscript*) utilizes a $3\times3\times16$ conv kernel.

3.2. Training Protocol

During the training phase, we train our framework with the following three stages. First, we pre-train the static saliency learning module (including ResNet-50 and PDC module) on the SOD datasets: *DUT-OMRON* [26], and two VSOD datasets: the training set of *DAVSOD* and *DAVIS* [16] dataset. We use the SGD optimizer with momentum rate set to 0.9 and weight decay set to 0.0005. The initial learning rate is set to 10^{-8} and is reduced with a factor of 0.1 in each epoch. We train the module for 2 epochs. The batch size is set to 2. Second, we train the whole model (SSAV) using the training set of *DAVSOD* and *DAVIS* [16] dataset. In this phase, we start with a learning rate of 10^{-6} and train the module for 2 epochs, reducing the learning rate in the 1-th and 2-th epoch with a factor 0.1. Finally, we fix the weights of the static saliency learning module and fine-tune the SSLSTM module using the training set of *DAVSOD* and *DAVIS* [16] dataset. For *DAVIS* [16] dataset, we use the object-level ground truths to train our module implicitly. For the proposed *DAVSOD* dataset, we explicitly use both object-level masks and fixation points as our training data. We follow the same data augmentation setting as [18], and set the sampling step set with $\{1, 2, 3, 4\}$.

Our *SSAV* model is implemented under the Caffe scientific computing platform. We use the CUDA toolkits and cuDNN accelerated lib in our implementation for high-performance GUP acceleration. The total training time is about 50 hours on only single NVIDIA TitanX GPU (12G Memory).

4. Metrics

4.1. Evaluation Metrics

MAE $\mathcal{M}$. We follow Perazzi *et al*. [15] to evaluate the *mean absolute error* (MAE) between a real-valued saliency map S and a binary ground-truth G for all image pixels:

$$\text{MAE} = \frac{1}{N}|S - G|, \tag{1}$$

where N is the total number of pixels. The MAE estimates the approximation degree between the saliency map and the ground-truth map, and it is normalized to $[0, 1]$. The MAE provides a direct estimate of conformity between estimated and ground-truth maps. However, for the MAE metric, small objects naturally assign a smaller error and larger objects have larger errors. The metric also can not tell where the error occurs [20].

Dataset	Year	Pub	#Vi.	#AF.	HQ	SIZE	AS	FP	EF	IL	DE	AN	DL
SegV1 [20]	2010	BMVC	6	244								✓	✓
SegV2 [8]	2013	ICCV	14	1, 065								✓	✓
FBMS [14]	2014	TPAMI	59	720	✓								
ViSal [23]	2015	TIP	17	193									
MCL [7]	2015	TIP	9	463								✓	
DAVIS [16]	2016	CVPR	50	3, 455	✓							✓	✓
UVSD [12]	2017	TCSVT	18	3, 262	✓								✓
VOS [10]	2018	TIP	200	7, 467		✓		✓					
DAVSOD	2019		**226**	**23, 938**	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: **Statistics of previous VSOD datasets and the proposed *DAVSOD* dataset.** *From left to right*: number of videos (**#Vi.**), number of annotated frames (**#AF.**), high quality annotation for dataset (**HQ**), large ($\geq$ 100 videos) size of the dataset (**SIZE**), whether consider attention shift (**AS**) phenomena, whether annotate salient objects according to human fixation points (**FP**), whether offer the eye fixation points of annotated salient objects (**EF**), whether provide instance-level annotation (**IL**), whether provide video description (**DE**), whether provide attribute annotation (**AN**), densely (per-frame) labeling (**DL**). Our dataset is the only one meeting all requirements. SegV1, and SegV2 originally introduced to evaluate tracking algorithms and then widely used for video segmentation and VSOD. SegV1 is a subset of SegV2. For the video clip of *penguins* in SegV2, only several penguins in the center are annotated as salient objects in the original ground truths, thus Liu *et al*. [12] relabeled this sequence to generate the higher quality annotations.

F-measure $\mathcal{F}$. F-measure is essentially a region based similarity metric. Following Cheng and Zhang *et al*. works [2, 28], we also provide the mean F-measure, max F-measure using varying fixed (0-255) thresholds, and F-measure using adaptive ([1]) threshold.

S-measure $\mathcal{S}$. Both MAE and F-measure metrics ignore the important structure information evaluation, whereas behavioral vision studies have shown that the human visual system is highly sensitive to structures in scenes [5]. Thus, we additionally include the structure metric S-measure [5] for a more comprehensive evaluation. The S-measure combines the region-aware ($\mathcal{S}_r$) and object-aware ($\mathcal{S}_o$) structural similarity as their final structure metric:

$$\mathcal{S} = \alpha * \mathcal{S}_o + (1 - \alpha) * \mathcal{S}_r, \tag{2}$$

where $\alpha \in [0, 1]$ is the balance parameter and set 0.5 as default.

4.2. Metric Statistics

For a given metric $\vartheta \in \{S_\alpha, F_\beta, mae, \zeta\}$ we consider different statistics. Let $V = \{S_i\}$ be the dataset of video sequence S_i. I_j^i denote the image of video sequence. Thus, $S_i = \{I_1^i, I_2^i, \ldots, I_j^i\}$. Let $\overline{\vartheta}(S_i)$ be the metric average on S_i. The *mean* is the average dataset statistic defined as $M_\vartheta(R) = \frac{1}{|R|}\sum_{S\in R}\overline{\vartheta}(S_i)$. Specifically, $\zeta_{dif}(R) = \frac{1}{|R|}\sum_{S\in R}|\overline{\zeta}_{gt}(S_i) - \overline{\zeta}_{sal}(S_i)|$. The average difference of metric over different datasets are summarized in Table 3.

4.3. More Experiment Results

In manuscript, we have shown that the proposed SSAV algorithm captures the saliency-shift phenomenon successfully. Here, we present more experiment results regarding to various of challenging scenes.

Given an input video sequence, the proposed SSAV model generates high-quality detection results of the salient object. Our model is capable of handling unconstrained videos that span a wide variety of situations including object with slow motion (bird in Fig. 8), occlusion (rabbit in Fig. 9), clutter (car in Fig. 10), shadow (ball in Fig. 11), simple background (cup in Fig. 12), non-ridge deformation (moving worm in Fig. 13), dynamic background (dancer in Fig. 14), and low contrast (train in Fig. 15).

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *IEEE CVPR*, pages 1597–1604, 2009. 1, 3

[2] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient Object Detection: A Benchmark. *IEEE TIP*, 24(12):5706–5722, 2015. 3

[3] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. *IEEE TIP*, 26(7):3156–3170, 2017. 1, 5, 10, 11, 12, 13

[4] Y. Chen, W. Zou, Y. Tang, X. Li, C. Xu, and N. Komodakis. Scom: Spatiotemporal constrained optimization for salient object detection. *IEEE TIP*, 27(7):3345–3357, 2018. 1, 5, 10, 11, 13

[5] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *IEEE ICCV*, pages 4548–4557, 2017. 1, 3

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 1

[7] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim. Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE TIP*, 24(8):2552–2564, 2015. 1, 3, 5, 7, 10, 11, 12, 13

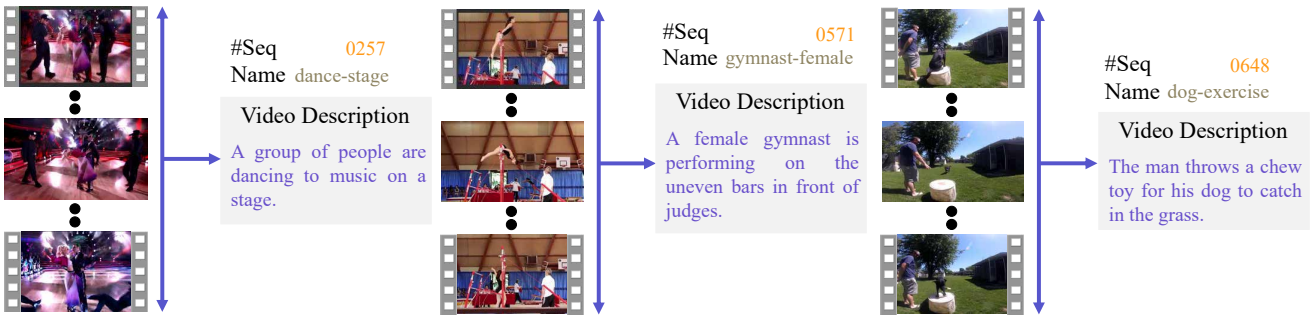


Figure 3: **Examples of video textual description in our *DAVSOD* dataset.**

#Seq	Name	Video description
0549	dance-break	The gymnast performs a double back flip in the gym
0552	bird-eagle	Two eagles are fighting in the wild
0554	bear	Two black bears are chasing each other around the tree
0556	whale-oceanarium	A killer whale cooperates very well under the keeper's instructions in the oceanarium
0566	woman-kitchen	A young woman is talking to the camera in the kitchen
0567	talking-street	A woman is playing a joke on the man by shaking his hand with dirty one
0570	dance-stage5	A man and a woman are dancing on the stage in front of the audience
0572	sheep-running	A group of sheep are running across the grass
0575	rhino	A little rhino walks behind an older rhino closely in the wild
0576	aeroplane-helicopter2	A helicopter lands in a clearing in the jungle
0577	car-bulldozer2	The man is teaching his daughter how to drive the bulldozer
0580	car-street3	The car was parked in the street waiting for the traffic lights
0590	dog-leopard	A dog is running after a leopard around the house
0593	color-biscuits	A woman is introducing the color biscuits while breaking one into halves
0600	fish-river	Fish are swimming freely in the clear water

Table 2: **Examples of sequence, name and the corresponding video description on the proposed *DAVSOD* dataset.** Please refer to the complete description in our website.

[8] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *IEEE ICCV*, pages 2192–2199, 2013. 1, 3, 5, 7, 13

[9] G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin. Flow guided recurrent neural encoder for video salient object detection. In *IEEE CVPR*, pages 3243–3252, 2018. 1, 5, 10, 11, 12, 13

[10] J. Li, C. Xia, and X. Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE TIP*, 27(1):349–364, 2018. 1, 3, 5, 7, 12

[11] S. Li, B. Seybold, A. Vorobyov, X. Lei, and C.-C. Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *ECCV*. Springer, 2018. 1, 5, 10, 11, 12, 13

[12] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. *IEEE TCSVT*, 27(12):2527–2542, 2017. 1, 3, 5, 10, 11, 12, 13

[13] Z. Liu, X. Zhang, S. Luo, and O. Le Meur. Superpixel-based spatiotemporal saliency detection. *IEEE TCSVT*, 24(9):1522–1540, 2014. 1, 5, 10, 11, 12, 13

[14] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *IEEE TPAMI*, 36(6):1187–1200, 2014. 1, 3, 5, 11

[15] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 1, 2

[16] 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 *IEEE CVPR*, pages 724–732, 2016. 1, 2, 3, 5, 11

[17] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä. Segmenting salient objects from images and videos. In *ECCV*, pages 366–379. Springer, 2010. 1, 5, 10, 11, 12, 13

[18] H. Song, W. Wang, S. Zhao, J. Sheng, and K.-M. Lam. Pyramid dilated deeper convLSTM for video salient object detection. In *ECCV*. Springer, 2018. 1, 2, 5, 10, 11, 12, 13

[19] Y. Tang, W. Zou, Z. Jin, Y. Chen, Y. Hua, and X. Li. Weakly supervised salient object detection with spatiotemporal cascade neural networks. *IEEE TCSVT*, 2018. 1, 5, 10, 11, 12, 13

[20] D. Tsai, M. Flagg, and J. Rehg. Motion coherent tracking with multi-label mrf optimization, algorithms. In *BMVC*, 2010. 2, 3, 7

	Metric	2010-2015 SIVM [17]	TIMP [29]	SPVM [13]	RWRV [7]	MB [27]	SAGM [22]	GFVM [23]	2016-2017 MSTM [21]	STBP [25]	SGSP [12]	SFLR [3]	2018 SCOM [4]†	SCNN [19]†	DLVS [24]†	FGRN [9]†	MBNM [11]†	PDBM [18]†	**SSAV**†
ViSal	max $\mathcal{F}\uparrow$	.522	.479	.700	.440	.692	.688	.683	.673	.622	.677	.779	.831	.831	.852	.848	.883	.888	*.939*
	ranking	16	17	9	18	10	11	12	14	15	13	8	6	6	4	5	3	2	1
	$\mathcal{S}\uparrow$	.606	.612	.724	.595	.726	.749	.757	.749	.629	.706	.814	.762	.847	.881	.861	.898	.907	.943
	ranking	17	16	13	18	12	10	9	10	15	14	7	8	6	4	5	3	2	1
	$\mathcal{M}\downarrow$	.197	.170	.133	.188	.129	.105	.107	.095	.163	.165	.062	.122	.071	.048	.045	.020	.032	.020
	ranking	18	16	13	17	12	9	10	8	14	15	6	11	7	5	4	1	3	1
FBMS-T	max $\mathcal{F}\uparrow$	.426	.456	.330	.336	.487	.564	.571	.500	.595	.630	.660	.797	.762	.759	.767	.816	.821	.865
	ranking	16	15	18	17	14	12	11	13	10	9	8	4	6	7	5	3	2	1
	$\mathcal{S}\uparrow$	.545	.576	.515	.521	.609	.659	.651	.613	.627	.661	.699	.794	.794	.794	.809	.857	.851	.879
	ranking	16	15	18	17	14	10	11	13	12	9	8	5	5	5	4	2	3	1
	$\mathcal{M}\downarrow$	.236	.192	.209	.242	.206	.161	.160	.177	.152	.172	.117	.079	.095	.091	.088	.047	.064	.040
	ranking	17	14	16	18	15	11	10	13	9	12	8	4	7	6	5	2	3	1
DAVIS-T	max $\mathcal{F}\uparrow$	.450	.488	.390	.345	.470	.515	.569	.429	.544	.655	.727	.783	.714	.708	.783	.861	.855	.861
	ranking	15	13	17	18	14	12	10	16	11	7	6	4	7	8	4	1	3	1
	$\mathcal{S}\uparrow$	.557	.593	.592	.556	.597	.676	.687	.583	.677	.692	.790	.832	.783	.794	.838	.887	.882	.893
	ranking	17	14	15	18	13	12	10	16	11	9	7	5	8	6	4	2	3	1
	$\mathcal{M}\downarrow$	.212	.172	.146	.199	.177	.103	.103	.165	.096	.138	.056	.048	.064	.061	.043	.031	.028	.028
	ranking	18	15	13	17	16	10	10	14	9	12	6	5	8	7	4	3	2	1
SegV2	max $\mathcal{F}\uparrow$	.581	.573	.618	.438	.554	.634	.592	.526	.640	.673	.745	.764	**	**	**	.716	.800	.801
	ranking	11	12	9	15	13	8	10	14	7	6	4	3				5	2	1
	$\mathcal{S}\uparrow$	.605	.644	.668	.583	.618	.719	.699	.643	.735	.681	.804	.815	**	**	**	.809	.864	.851
	ranking	14	11	10	15	13	7	8	12	6	9	5	3				4	1	2
	$\mathcal{M}\downarrow$	.251	.116	.108	.162	.146	.081	.091	.114	.061	.124	.037	.030	**	**	**	.026	.024	.023
	ranking	15	11	9	14	13	7	8	10	6	12	5	4				3	2	1
UVSD	max $\mathcal{F}\uparrow$	.293	.338	.404	.281	.339	.414	.426	.336	.403	.544	.562	.420	.550	.564	.630	.550	.863	.801
	ranking	17	15	12	18	14	11	9	16	13	8	5	10	6	4	3	6	1	2
	$\mathcal{S}\uparrow$	.481	.537	.581	.536	.563	.629	.628	.551	.614	.601	.713	.555	.712	.721	.745	.698	.901	.861
	ranking	18	16	12	17	13	8	9	15	10	11	5	14	6	4	3	7	1	2
	$\mathcal{M}\downarrow$	.260	.178	.146	.180	.169	.111	.106	.145	.105	.165	.059	.206	.075	.060	.042	.079	.018	.025
	ranking	18	15	12	16	14	10	9	11	8	13	4	17	6	5	3	7	1	2
MCL	max $\mathcal{F}\uparrow$	.420	.598	.595	.446	.261	.422	.406	.313	.607	.645	.669	.422	.628	.551	.625	.698	.798	.774
	ranking	15	9	10	12	18	13	16	17	8	5	4	13	6	11	7	3	1	2
	$\mathcal{S}\uparrow$	.548	.642	.665	.577	.539	.615	.613	.540	.700	.679	.734	.569	.730	.682	.709	.755	.856	.819
	ranking	16	11	10	14	18	12	13	17	7	9	4	15	5	8	6	3	1	2
	$\mathcal{M}\downarrow$	.185	.113	.105	.167	.178	.136	.132	.171	.078	.100	.054	.204	.054	.060	.044	.119	.021	.027
	ranking	17	10	9	14	16	13	12	15	7	8	4	18	4	6	3	11	1	2
VOS-T	max $\mathcal{F}\uparrow$	.439	.401	.351	.422	.562	.482	.506	.567	.526	.426	.546	.690	.609	.675	.669	.670	.742	.742
	ranking	14	17	18	16	9	13	12	8	11	15	10	3	7	4	6	5	1	1
	$\mathcal{S}\uparrow$	.558	.575	.511	.552	.661	.619	.615	.657	.576	.557	.624	.712	.704	.760	.715	.742	.818	.819
	ranking	15	14	18	17	8	11	12	9	13	16	10	6	7	3	5	4	2	1
	$\mathcal{M}\downarrow$	.217	.215	.223	.211	.158	.172	.162	.144	.163	.236	.145	.162	.109	.099	.097	.099	.078	.073
	ranking	16	15	17	14	9	13	10	7	12	18	8	10	6	4	3	4	2	1
DAVSOD-T	max $\mathcal{F}\uparrow$	.298	.395	.358	.283	.342	.370	.334	.344	.410	.426	.478	.464	.532	.521	.563	.510	.562	.630
	ranking	17	11	13	18	15	12	16	14	10	9	7	8	4	5	2	6	3	1
	$\mathcal{S}\uparrow$	.486	.563	.538	.504	.538	.565	.553	.532	.568	.577	.624	.599	.657	.637	.673	.654	.678	.699
	ranking	18	12	14	17	14	11	13	16	10	9	7	8	4	6	3	5	2	1
	$\mathcal{M}\downarrow$	.288	.195	.202	.245	.228	.184	.167	.211	.160	.207	.143	.220	.139	.140	.109	.170	.127	.098
	ranking	18	11	12	17	16	10	8	14	7	13	6	15	4	5	2	9	3	1
overall ranking		18	16	14	17	15	10	11	13	9	12	7	8	6	5	3	4	2	1
Run time		72.4s	69.2s	56.1s	18.3s	0.02s	45.4s	53.7s	0.02s	49.49s	51.7s	119.4s	38.8s	38.5s	0.47s	0.09s	2.63s	0.05s	0.049s
ranking		17	16	15	8	1	11	14	1	12	13	18	10	9	6	5	7	4	3

Table 3: **Benchmarking results of 17 state-of-the-art VSOD models on 7 datasets**: *SegV2* [8], *FBMS* [14], *ViSal* [23], *MCL* [7], *DAVIS* [16], *UVSD* [12], *VOS* [10] and the proposed *DAVSOD*. Note that TIMP was only tested on 9 short sequences of *VOS* because it cannot handle long videos. "**" indicates the model has been trained on this dataset. "-T" indicates the results on the test set of this dataset. "†" indicates deep learning model. "score" indicates the GPU time. The top three models are marked in red, blue, and green color, respectively.

[21] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien. Real-time salient object detection with a minimum spanning tree. In *IEEE CVPR*, pages 2334–2342, 2016. 1, 5, 10, 11, 12, 13

[22] W. Wang, J. Shen, and F. Porikli. Saliency-aware geodesic video object segmentation. In *IEEE CVPR*, pages 3395–3402, 2015. 1, 5, 10, 11, 12, 13

[23] W. Wang, J. Shen, and L. Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE TIP*, 24(11):4185–4196, 2015. 1, 3, 5, 7, 10, 11, 12, 13

[24] W. Wang, J. Shen, and L. Shao. Video salient object detection via fully convolutional networks. *IEEE TIP*, 27(1):38–49, 2018. 1, 5, 10, 11, 12, 13

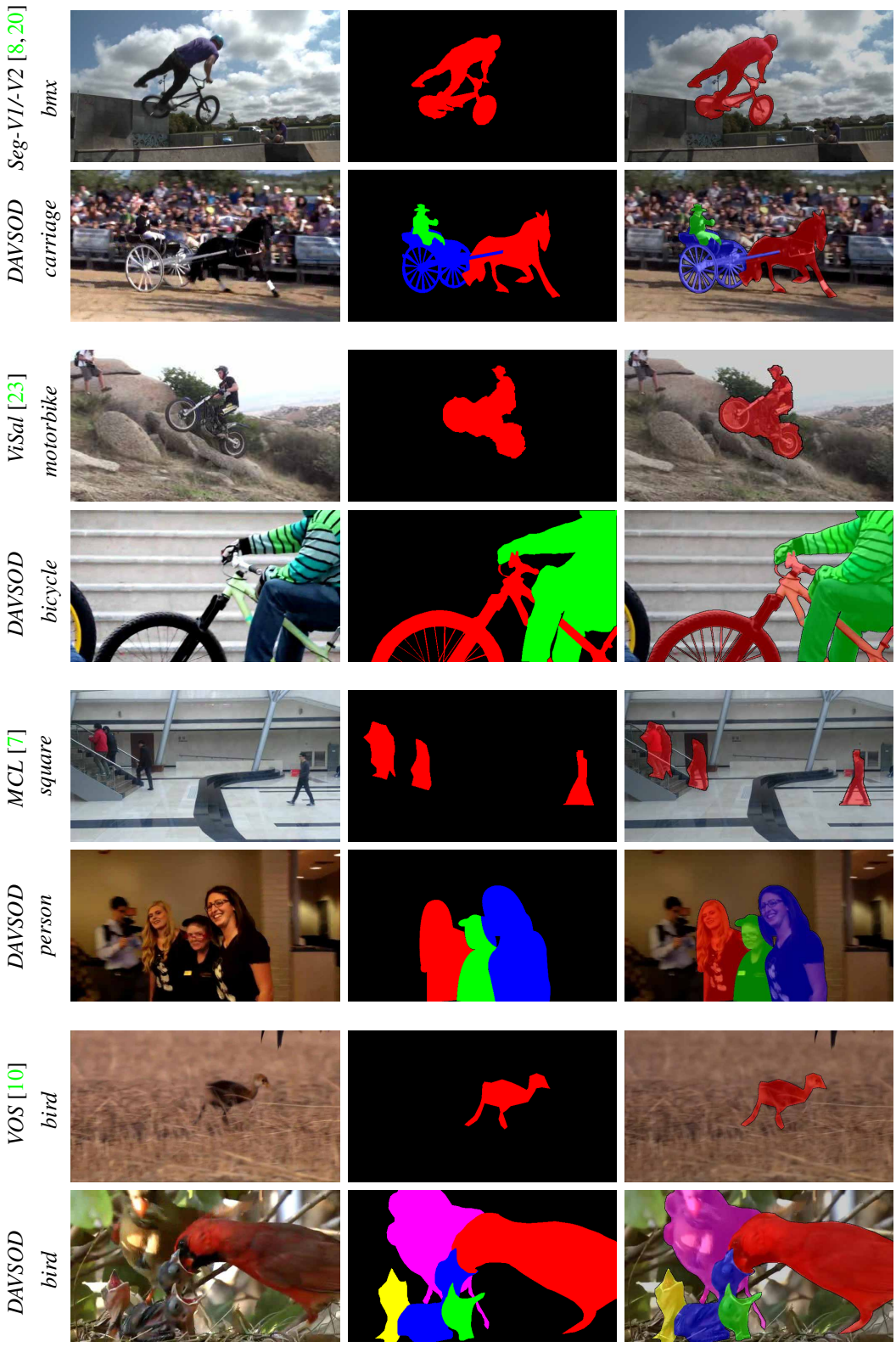
[25] T. Xi, W. Zhao, H. Wang, and W. Lin. Salient object detection with spatiotemporal background priors for video. *IEEE TIP*, 26(7):3425–3436, 2017. 1, 5, 10, 11, 12, 13

[26] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *IEEE CVPR*, pages 3166–3173, 2013. 2

[27] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech. Minimum barrier salient object detection at 80 fps. In *IEEE ICCV*, pages 1404–1412, 2015. 1, 5, 10, 11, 12, 13

[28] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *IEEE ICCV*, pages 202–211, 2017. 3

[29] F. Zhou, S. Bing Kang, and M. F. Cohen. Time-mapping using space-time saliency. In *IEEE CVPR*, pages 3358–3365, 2014. 1, 5, 10, 11, 13



(a) Image (b) GT (c) Overlap

Figure 4: **Compared with previous object-level datasets** which are labeled with coarse boundary (the spoke in "*bmx*" sequence, motorbike in 3^{rd} row) or polygons (*e.g.*, person in *MCL*, bird in *VOS*). However, the proposed object-/instance-level *DAVSOD* dataset is labeled with smooth fine boundaries (*e.g.*, the wheel of carriage in 1^{st} row, the spoke of bicycle in 4^{th} row, bird in last row, *etc.*).

Figure 5: **Sample sequences from our dataset**, with instance-level ground truth segmentation masks and fixation map overlayed. Please refer to the accompanying video (VisualSaliency.mp4) for a visualization of the dataset.

Figure 6: **Sample sequences from our dataset**, with instance-level ground truth segmentation masks and fixation map overlayed. Please refer to the accompanying video (VisualSaliency.mp4) for a visualization of the dataset.

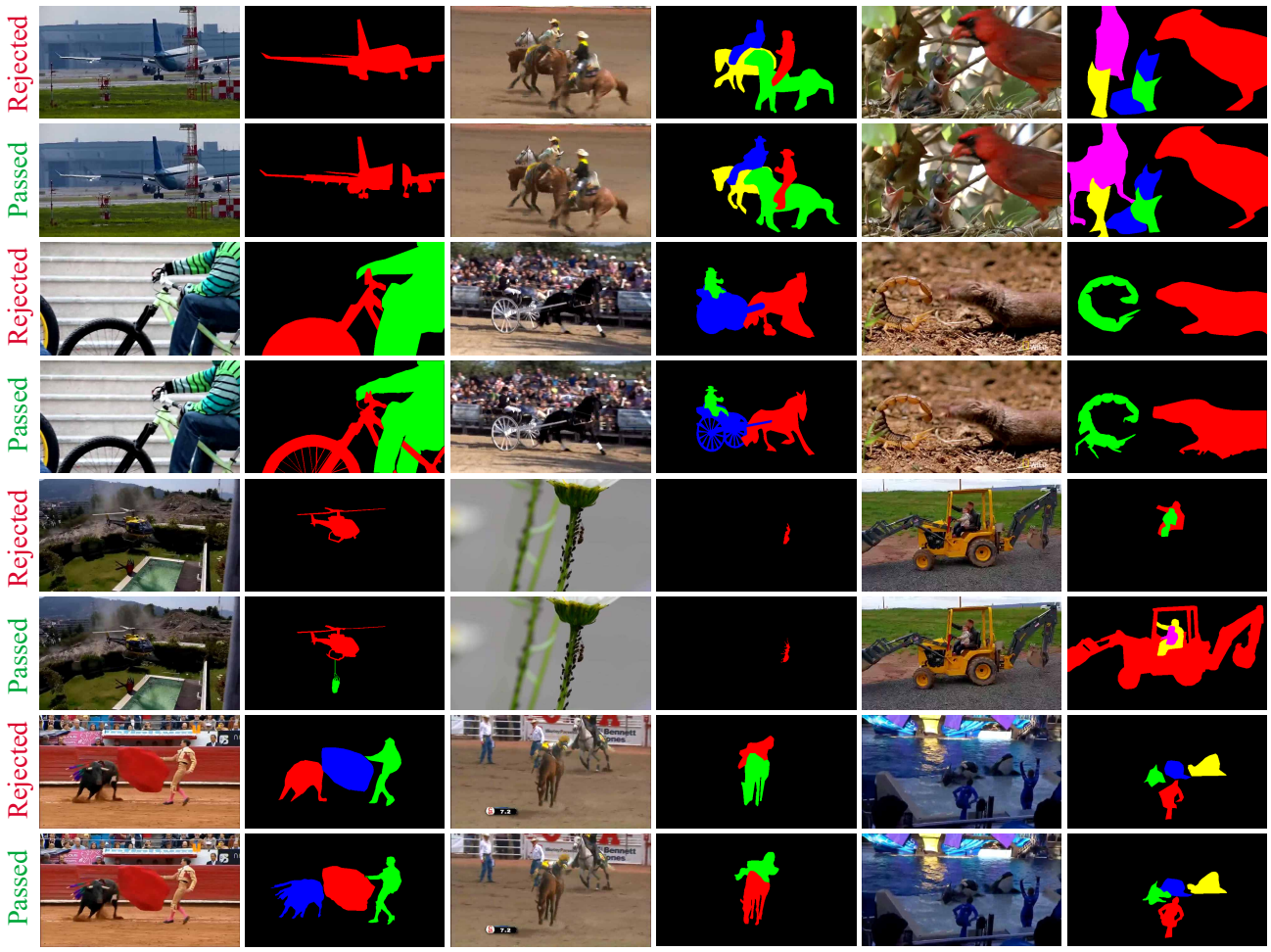


Figure 7: **Examples of cases that passed or were rejected segmentation in the verification stage.** Zoom-in for details.

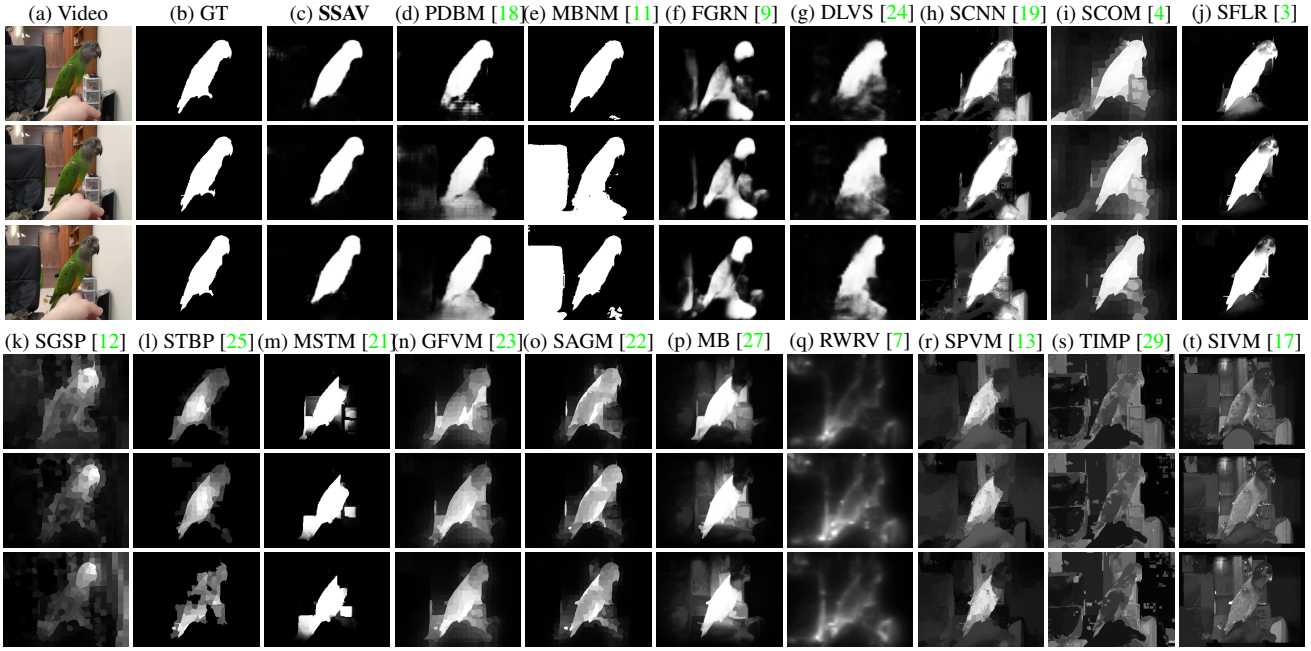


Figure 8: **Qualitative comparison against other top-performing VSOD models.** The *bird* sequence from the ViSal [23] dataset.

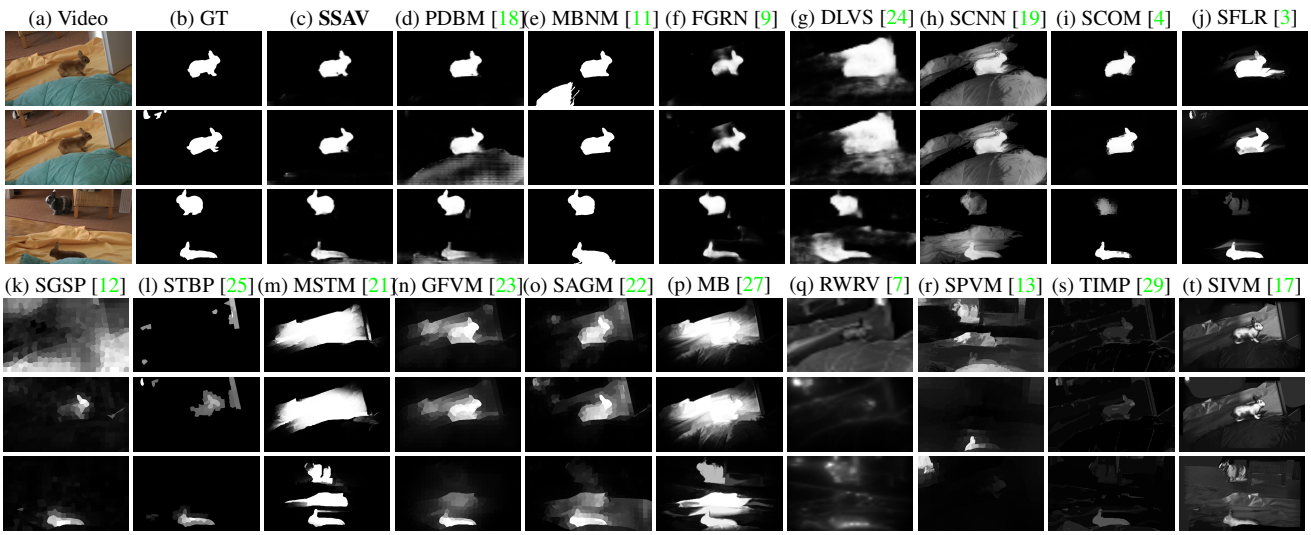


Figure 9: **Qualitative comparison against other top-performing VSOD models.** The *rabbits03* sequence from the FBMS [14] dataset. Zoom-in for details.

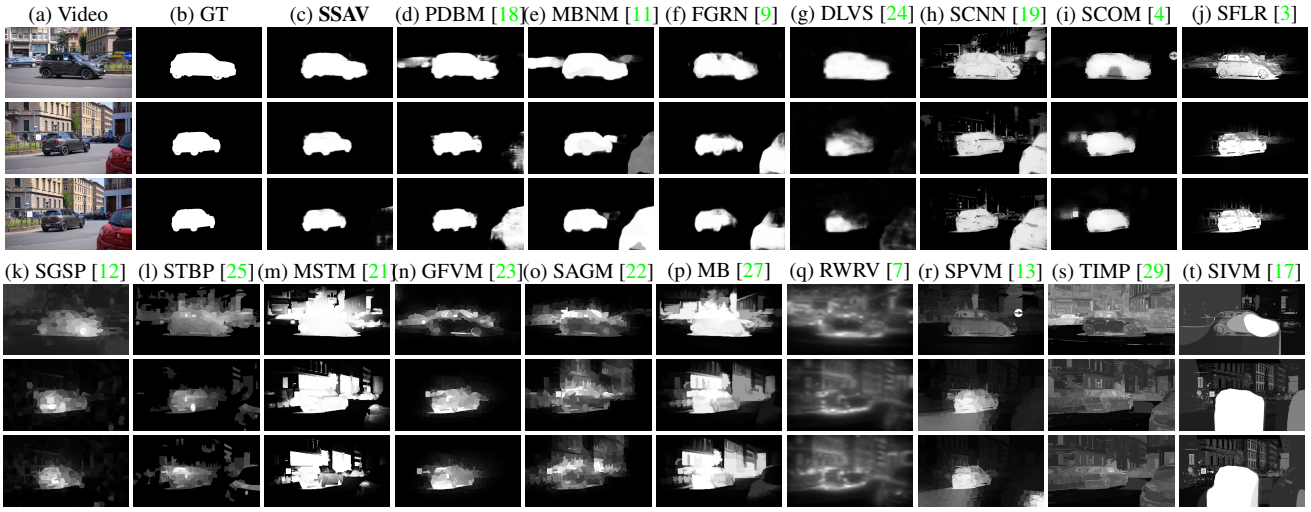


Figure 10: **Qualitative comparison against other top-performing VSOD models.** The *car-roundabout* sequence from the DAVIS [16] dataset. Zoom-in for details.

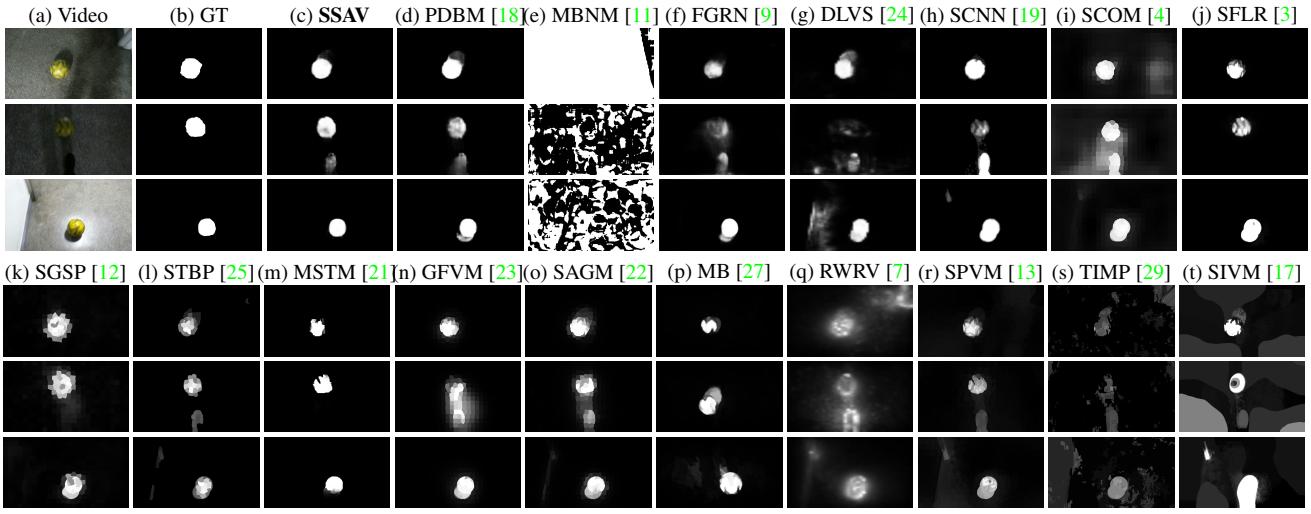


Figure 11: **Qualitative comparison against other top-performing VSOD models.** The *Ball* sequence from the MCL [7] dataset.

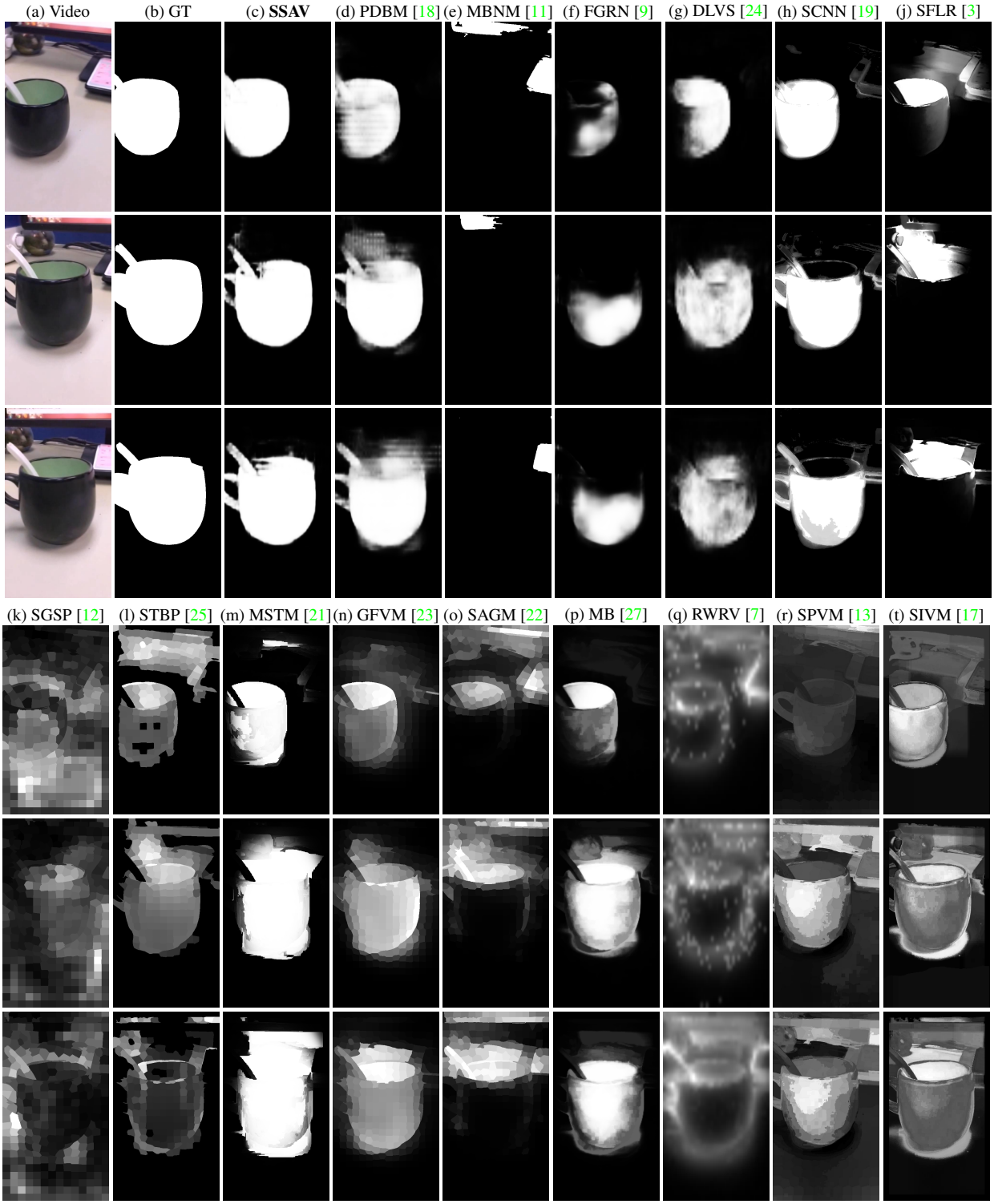


Figure 12: **Qualitative comparison against other top-performing VSOD models.** The *cup* sequence from the VOS [10] dataset. Zoom-in for details.

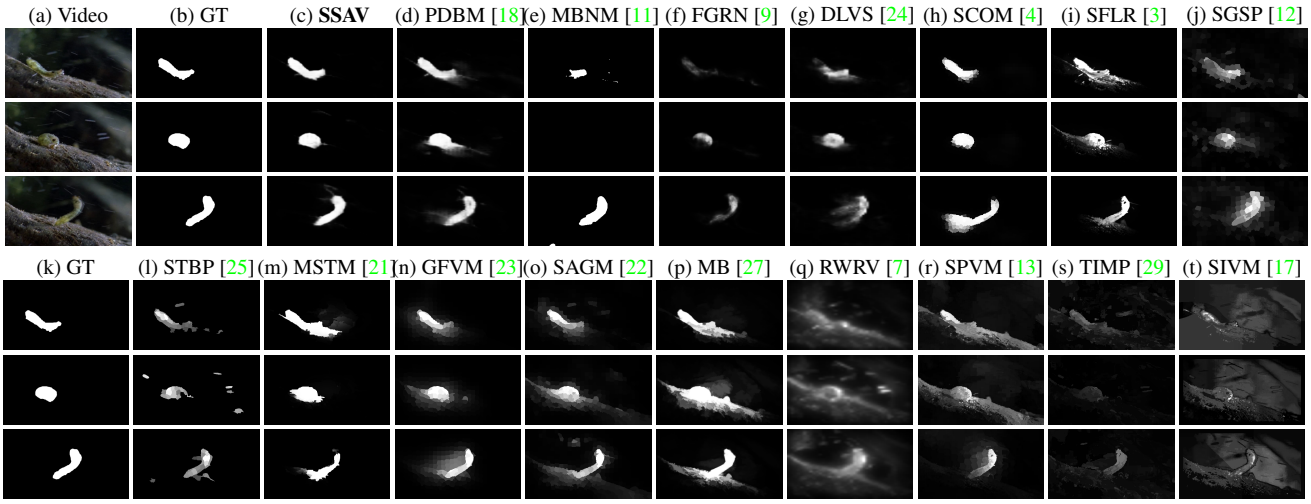


Figure 13: **Qualitative comparison against other top-performing VSOD models.** The *worm* sequence from the SegTrack-V2 [8] dataset. Zoom-in for details.

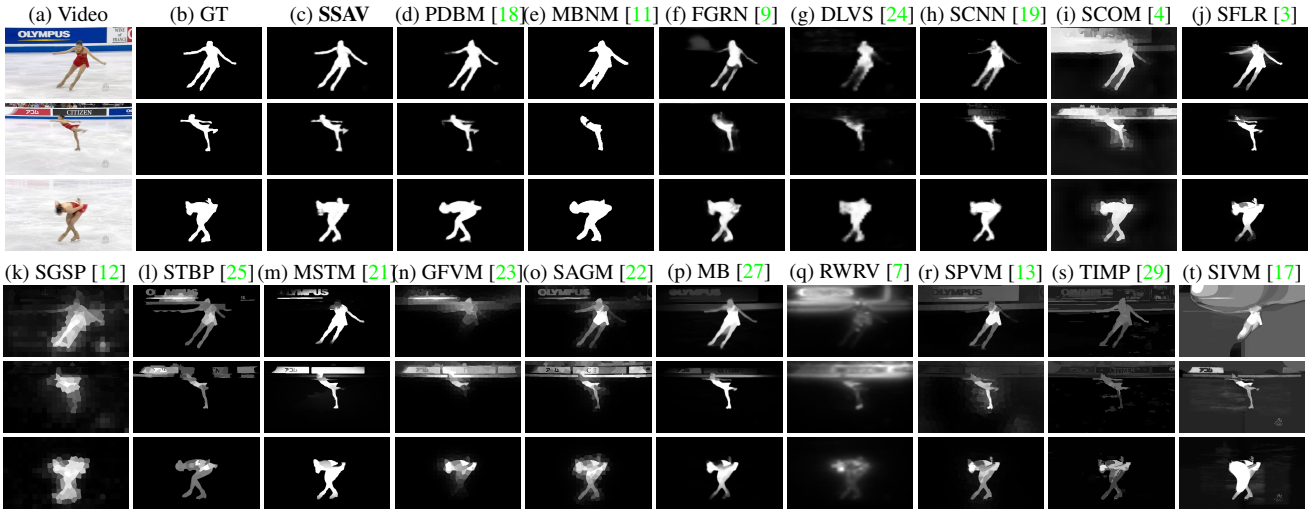


Figure 14: **Qualitative comparison against other top-performing VSOD models.** The *yunakim_long2* sequence from the UVSD [12] dataset. Zoom-in for details.

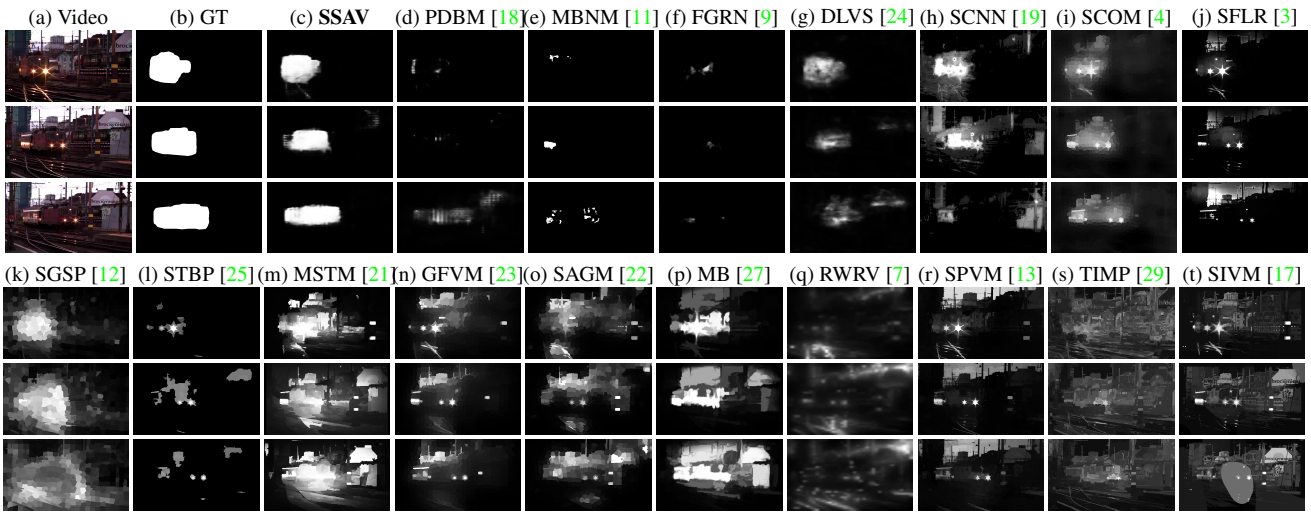


Figure 15: **Qualitative comparison against other top-performing VSOD models.** The *train* sequence from the proposed DAVSOD dataset. Zoom-in for details.