DVSOD: RGB-D Video Salient Object Detection

Jingjing Li[1], **Wei Ji**[1,2*], **Size Wang**[1], **Wenbo Li**[2], **Li Cheng**[1]
[1]University of Alberta, Canada
[2]Samsung Research America AI Center, USA
{jingjin1, wji3, size1, lcheng5}@ualberta.ca, wenbo.li1@samsung.com

Abstract

Salient object detection (SOD) aims to identify standout elements in a scene, with recent advancements primarily focused on integrating depth data (RGB-D) or temporal data from videos to enhance SOD in complex scenes. However, the unison of two types of crucial information remains largely underexplored due to data constraints. To bridge this gap, we in this work introduce the DViSal dataset, fueling further research in the emerging field of RGB-D video salient object detection (**DVSOD**). Our dataset features 237 diverse RGB-D videos alongside comprehensive annotations, including object and instance-level markings, as well as bounding boxes and scribbles. These resources enable a broad scope for potential research directions. We also conduct benchmarking experiments using various SOD models, affirming the efficacy of multimodal video input for salient object detection. Lastly, we highlight some intriguing findings and promising future research avenues. To foster growth in this field, *our dataset and benchmark results are publicly accessible at:* `https://dvsod.github.io/`.

1 Introduction

Salient object detection (SOD) is a fascinating field of study that aims at identifying and distinguishing the most eye-catching components in static images [59, 91] or dynamic videos [66, 89]. This intriguing concept hails from cognitive science research into human visual attention behavior - our incredible ability to swiftly direct attention to the most valuable portions of visual scenes [4]. As such, SOD assumes an indispensable role in a myriad of real-world applications [52, 53, 30, 50, 69, 96, 24, 93, 81], such as image retrieval, video editing/analysis, medical diagnosis, object detection, and target tracking.

This field has witnessed significant advances in recent years, particularly in RGB-image based SOD [65, 90, 82, 72]. However, it has been observed that performance often degrades when objects and their surroundings share similar appearances or when backgrounds are heavily cluttered. With the growing ubiquity of 3D imaging sensors in depth cameras, such as Kinect and Intel RealSense, as well as in mobile devices like iPhone 13 Pro, Huawei Mate 50, and Samsung Galaxy S21, the value of fully leveraging RGB-D information for SOD has gained considerable research interest [14, 58, 94, 95]. As visualized in Fig. 1, paired depth maps contribute valuable 3D geometric information that helps the discernment of object silhouettes against backgrounds. Consequently, augmenting RGB images with depth maps as additional input dramatically enhances the accuracy in localizing salient objects in complex scenes, *i.e.*, '*w*/ RGB' *vs.* '*w*/ RGBD' in Fig. 1.

On another front, considering the dynamic nature of real-world scenes, video salient object detection (VSOD) has been a prominent focus of research [19, 56, 71]. In contrast to static images, dynamic video scenario presents considerable difficulties due to the diversity of motion patterns, occlusions, blurring, large object deformations, and the inherent complexity of human visual attention behavior.

[*] Corresponding author.

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

https://doi.org/10.52202/075280-0383

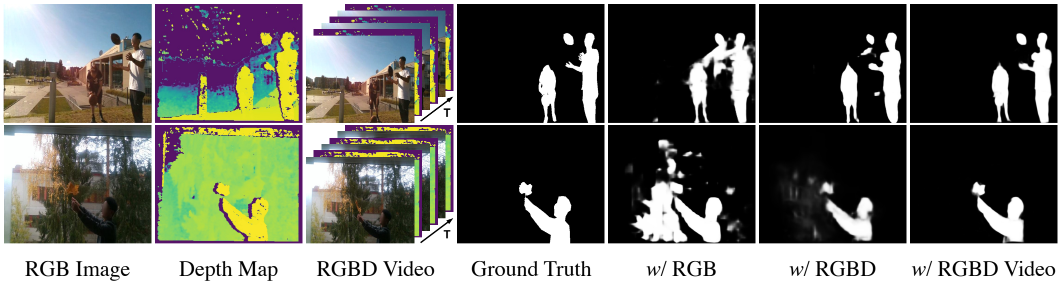


Figure 1: Visual illustration of the advantages of employing RGBD videos for salient object detection. The last three columns exhibit the segmentation results achieved using different input modalities.

By leveraging temporal contexts within a video sequence, VSOD methods [19, 63] have demonstrated improved capture of moving foreground objects and modeling of visual attention shifts in dynamic environments.

While there have been significant strides in both RGB-D SOD and VSOD individually, there remains a paucity of research into combining multimodal and temporal contexts, both of which are vital for accurately detecting the salient objects in a scene. This lack is primarily due to the absence of suitable benchmark datasets. As an alternative, Lu *et al.* [48] have made the initial attempt to amalgamate existing VSOD datasets with estimated depth maps. However, these estimated depth maps often fall short of accurately reflecting the real world, and are prone to errors introduced by depth estimation methods, compared to data captured by actual depth sensors.

To bridge the existing gap and prosper the progress in RGB-D video SOD (*i.e.*, DVSOD), we introduce the first real-world RGB-**D Vi**deo **Sal**iency detection dataset, abbreviated as DViSal, in this study. Specifically, the DViSal dataset comprises 237 RGB-D videos at a frame rate of 25 f/s, including 175,442 RGB-D pairs in total and 7,117 annotated frames. Besides, the videos were carefully selected to encompass a broad diversity of real scenarios and motion patterns. Another distinguishing characteristic of the DViSal dataset is the variety of annotation formats provided. In addition to the conventional object-level annotations used in SOD tasks, the dataset offers more comprehensive instance-level annotations, as well as weak annotations consisting of bounding boxes and scribbles. These diverse annotation sets may serve as valuable resources for a wider range of potential research directions, such as instance-level DVSOD, weakly-supervised DVSOD, and more.

We further benchmark the DVSOD by implementing 11 cutting-edge SOD models on the newly-proposed DViSal dataset, including *RGB-image-based* SOD models, *RGB-D* SOD models, and *VSOD* models. Additionally, we have assembled a straightforward DVSOD model to dissect the specific contributions offered by depth maps and video contexts. The empirical results underscore the utility of both the multimodal RGB-D cues and temporal video contexts in accurately identifying salient objects. In particular, the DVSOD model is shown to significantly enhance the performance of the base RGB network [73], as evidenced by a noteworthy increase in the F-measure (*i.e.*, 0.548 to 0.610). Lastly, we contemplate and highlight promising avenues for future research. We observe that the field of DVSOD is far from fully explored, with substantial room for improvement and innovation. This research presents the community with an exciting opportunity to delve deeper into this burgeoning field.

2 Related Works

In the era of deep learning, benchmark datasets have become the foundational infrastructure for advancing the state-of-the-art in computer vision research. Prominent publicly available benchmarks, such as DUTS [70], SOC [17], and PASCAL-S [42], have greatly contributed to the remarkable progress in RGB-image-based salient object detection (SOD) [73, 65, 72] over the past decade. Despite their significant contributions, these benchmarks primarily center their attention on RGB images, overlooking the valuable depth data that can enhance the discrimination of object silhouettes from backgrounds.

With the rising popularity of 3D imaging sensors, acquiring RGB-D data has become easier with modern equipments (*e.g.*, Kinect, Samsung Galaxy S21 and iPhone 13 Pro). This development has

https://doi.org/10.52202/075280-0383

Table 1: Overview of existing RGB SOD, RGB-D SOD, VSOD datasets, and our proposed DViSal dataset. The table shows whether each dataset provides depth maps/video sequences (Depth/Video), object-level ground truths (Obj-GT), instance-level annotations (Ins-GT), and weak supervision signals (Weak-GT) such as bounding boxes or scribbles. The table also lists the number of videos (#Video), annotations (#GT), and instances (#Ins) each dataset includes.

Dataset	Year	Depth	Video	Obj-GT	Ins-GT	Weak-GT	#Video	#GT	#Ins
MSRA-B [46]	2007	✗	✗	✓	✗	✓	-	5,000	-
SED1&2 [2]	2007	✗	✗	✓	✗	✗	-	200	-
ASD [1]	2009	✗	✗	✓	✗	✗	-	1,000	-
SOD [51]	2010	✗	✗	✓	✗	✗	-	300	-
M10K [8]	2011	✗	✗	✓	✗	✗	-	10,000	-
DUT-O [76]	2013	✗	✗	✓	✗	✓	-	5,168	-
ECSSD [74]	2013	✗	✗	✓	✗	✗	-	1,000	-
PASCAL-S [42]	2014	✗	✗	✓	✗	✗	-	850	-
HKU-IS [36]	2015	✗	✗	✓	✗	✗	-	4,447	-
SOS [78]	2015	✗	✗	✗	✗	✓	-	6,900	-
MSO [78]	2015	✗	✗	✗	✗	✓	-	1,224	-
DUTS [70]	2017	✗	✗	✓	✗	✓	-	15,572	-
SOC [17]	2022	✗	✗	✓	✓	✓	-	6,000	5,776
STERE [55]	2012	✓	✗	✓	✗	✗	-	1,000	-
NJU2K [31]	2014	✓	✗	✓	✗	✗	-	1,985	-
NLPR [60]	2014	✓	✗	✓	✗	✗	-	1,000	-
DES [9]	2014	✓	✗	✓	✗	✗	-	135	-
LFSD [41]	2014	✓	✗	✓	✗	✗	-	100	-
SSD [97]	2017	✓	✗	✓	✗	✗	-	80	-
DUT-D [62]	2019	✓	✗	✓	✗	✗	-	1,200	-
SIP [14]	2020	✓	✗	✓	✓	✗	-	929	1,471
ReDWeb-S [44]	2021	✓	✗	✓	✗	✗	-	3,179	-
COME [79]	2021	✓	✗	✓	✓	✓	-	15,625	26,862
SegV2 [35]	2013	✗	✓	✓	✗	✗	14	1,065	-
FBMS [56]	2014	✗	✓	✓	✗	✗	59	720	-
MCL [32]	2015	✗	✓	✓	✗	✗	9	463	-
ViSal [71]	2015	✗	✓	✓	✗	✗	17	193	-
DAVIS [61]	2016	✗	✓	✓	✗	✗	50	3,455	-
UVSD [47]	2017	✗	✓	✓	✗	✗	18	3,262	-
VOS [37]	2018	✗	✓	✓	✗	✗	200	7,467	-
DAVSOD [15]	2019	✗	✓	✓	✓	✗	226	23,938	39,498
Our DViSal	-	✓	✓	✓	✓	✓	237	7,117	20,226

spurred the creation of several RGB-D datasets aimed at addressing the segmentation of salient objects in complex scenarios. For instance, the STERE dataset [55], which serves as the first collection of stereoscopic photos in this domain, boasts of 797 samples. The NJU2K dataset [31], on the other hand, houses 1,985 paired images in its most recent version, sourced from various platforms such as the internet, 3D movies, and images captured by a Fuji W3 stereo camera. Large-scale benchmarks such as ReDWeb-S [44] and COME [79] have been curated to validate the scalability of diverse models, thereby fostering advancement in this domain. Many novel RGB-D models [28, 34, 23, 16, 85, 86, 45, 38, 39] have been developed, showing that depth data, which includes spatial structure and 3D layout cues within a scene, can assist segmentation in challenging scenarios. For example, [34] proposed a prototype sampling network to adaptively weight RGB and depth superpixels according to a reliance selection module. [45] explored visual transformer to capture global contexts, achieving promising results.

It is worth noting that the existing RGB-D SOD datasets rely on single static images. The lack of a mechanism to account for the temporal contexts might impede their effectiveness when dealing with video inputs that depict dynamic scenes, which are omnipresent in our daily lives [87, 83, 40, 26]. As depicted in Table 1, the current SOD datasets provide as input either single pairs of RGB and depth images (*e.g.*, STERE [55] and NJU2K [31]), or RGB only video sequences (*e.g.*, VOS [37] and DAVSOD [15]). There unfortunately lacks a suitable dataset to provide both types of critical information - spatial layout cues derived from depth data and temporal context derived from dynamic video sequences, in addition to the RGB images.

In this work, we focus on a systematic exploration of RGB-D video salient object detection to bridge this existing gap. We introduce a carefully curated benchmark dataset named DViSal. Table 1

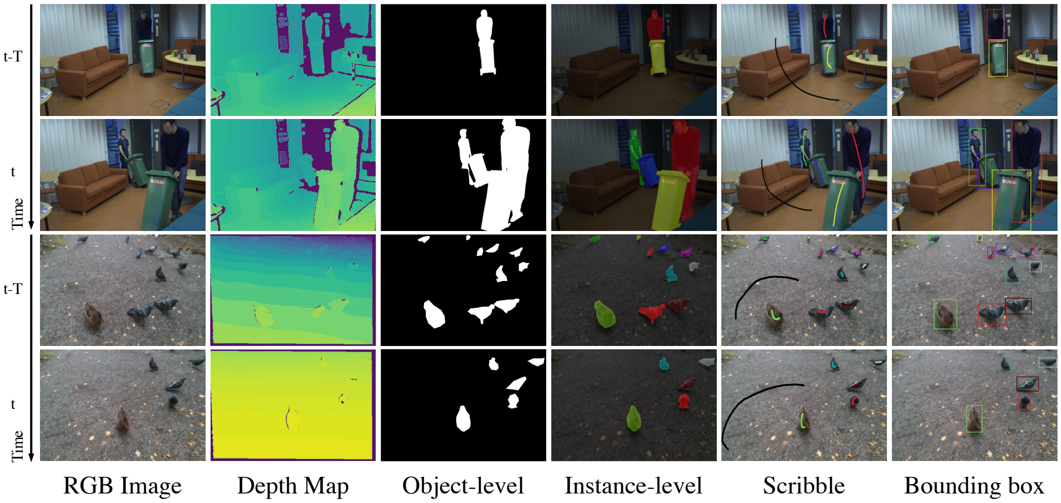


Figure 2: Examples of the DViSal dataset. We provide diverse annotations, including fully-supervised object-/instance-level markings, as well as weakly-supervised scribbles and bounding boxes.

presents the statistics of our newly launched DViSal alongside its counterparts. In addition, we undertake extensive benchmarking experiments that verify the advantages of utilizing RGB-D videos for saliency detection.

3 The DVSOD Dataset

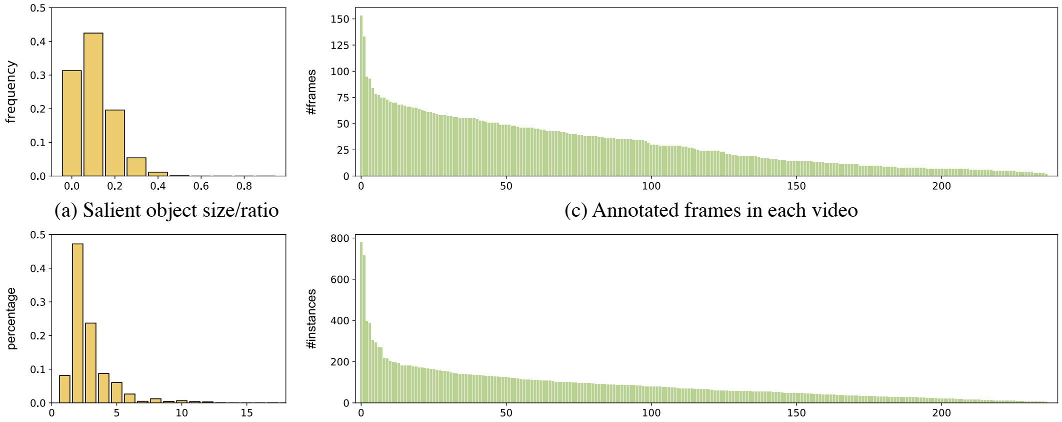
In this section, we introduce the new RGB-**D Vi**deo **Sal**iency detection dataset, *i.e.*, ***DViSal***. We first describe the construction process of the DViSal dataset, and then analyze its statistical results.

3.1 Dataset Construction

Dataset Collection. The main principle of data acquisition is to provide a comprehensive collection of calibrated RGB and depth video sequences, covering a broad spectrum of scenarios. Toward this objective, we initially gathered over 703 RGB-D videos, sourced from multiple repositories, including CDTB [49], People [68], PTB [67], Scene [33], DET [75], Tracklam [5], and Track3D [77]. The collected video samples exhibit diverse locations (*e.g.*, parks, campuses, streets and indoor scenes), and are recorded under a wide range of challenging conditions (*e.g.*, cluttered backgrounds, low-light conditions, and reflective environments). We are committed to ensuring the quality of our dataset. To this end, we meticulously remove unqualified videos or frames - ones that are blurry, misaligned, or feature ambiguous salient objects. After this selection process, we consolidate the DViSal dataset, which consists of 237 high-quality RGB-D videos, with 175,442 paired frames in total. Some visual examples from our dataset are illustrated in Fig. 2.

Dataset Annotation. Diverging from most established SOD datasets that primarily offer object-level saliency masks, our proposed dataset aims to deliver more comprehensive annotations. We employ the Labelme toolkit to annotate the collated RGB-D videos. The acquisition of the ground-truth saliency adheres to the annotation principle utilized in widely-recognized SOD datasets [62, 15, 41]. Initially, five annotators select candidate salient objects based on their initial instinct. Subsequently, a majority voting strategy is deployed to finalize the salient objects. Nonetheless, annotating a large-scale DViSal dataset presents additional challenges compared to labeling an RGB image dataset. First, our DViSal dataset includes many challenging scenes recorded under diverse conditions, *e.g.*, multiple objects, transparent objects, similar foreground and background, and cluttered environment. These complexities make it more difficult to identify whole objects and distinguish their silhouettes. Second, for video sequences, it becomes necessary to account for the attention shift [15] of salient objects within a single video. For example, an object labeled as "background" in one frame may be categorized as "salient" in another if the camera view changes. To overcome these challenges, we put into action several proactive measures. We carry out a visualization assistance process that overlays depth heatmaps onto corresponding RGB images, facilitating easier verification of salient objects in complex scenarios by the annotators. Moreover, for videos showcasing saliency shifting, two additional inspectors are required to review the initial annotations on an item-by-item basis, identifying any mislabeled samples and sending them back to the annotators for corrections and re-verification. Through these efforts, we successfully obtain a large-scale DViSal dataset comprising 7,117 high-quality annotations. These include object-level saliency ground truths, instance-level IDs, and weak supervision signals, such as bounding boxes and scribbles.

Dataset Splits. The entire dataset is partitioned into training, validation and test sets, which consist of 103, 26, and 108 videos, respectively, with 3,560, 200, and 3,357 annotated frames each. To be specific, the training set includes the entire CDTB [49] dataset, which contains 71 video sequences, and an additional 32 randomly selected videos from the PTB [67], Tracklam [5] and Track3D [77] datasets. The validation set is made up of 26 videos sourced from PTB, serving the purpose of model performance assessment during training. The remaining 108 videos, extracted from Tracklam, Track3D, DET [75], People [68], and Scene [33] constitute the test set, designed to evaluate the performance of the models. The videos derived from the DET, People, and Scene datasets are fully preserved within the test set, facilitating a comprehensive evaluation of various SOD models and corroborating their ability to generalize across different scenarios. As presented in Table 2, we report the benchmarking results for each subset and provide an overall measurement.



(a) Salient object size/ratio

(c) Annotated frames in each video

(b) Salient object numbers

(d) Annotated salient object instances in each video

Figure 3: Statistics of the proposed DViSal dataset in terms of (a) salient object size, (b) salient object numbers at image level, as well as (c) the number of annotated frames and (d) instances at video level.

3.2 Dataset Statistics

Comparison with Existing Datasets. Table 1 presents a comparative overview of popular SOD datasets and our proposed DViSal dataset. As observed, our DViSal dataset offers several notable advantages over existing datasets. First, DViSal integrates more rich data types in both RGB and depth views and video sequences. This stands in contrast to current datasets such as COME [79] and DAVSOD [15], which either provide single pairs of RGB and depth images or solely RGB video sequences as input. Second, the DViSal furnishes a more comprehensive set of annotations. Compared to typical SOD datasets that often involve only one type of object-level annotation [31, 36, 60, 62, 74], the diverse annotation sets in our DViSal dataset enable the exploration of additional domains such as instance-level or weakly-supervised DVSOD. This expanded capacity for research could foster a wider spectrum of research opportunities and catalyze the development of more sophisticated algorithms. Additionally, our dataset covers a wide range of scenarios and contains many challenging cases, such as cluttered background in Fig. 1 and multi-object scenario in Fig. 2. The diversity will be a valuable asset in assessing the scalability of saliency models.

Statistical Analysis. Fig. 3 provides the statistical results of our DViSal dataset from several dimensions. (1) *Size of salient objects*: The sizes of salient objects in the dataset significantly impact the performance of saliency detection, as smaller objects are generally more challenging to detect. In our analysis, we define the object size (OS) as the proportion of salient pixels to the total number of pixels in the annotated frames. The results in Fig. 3 (a) reveal that the DViSal dataset exhibits a relatively small proportion of images with large objects, and the size of salient objects varies widely

https://doi.org/10.52202/075280-0383

within a broad range of 0.02% to 54.55%. Moreover, the mean OS score is 14.65%, which is notably appealing compared to the score of 23.16% in current popular RGB-D datasets as reported in [29]. (2) *Number of salient objects*: Fig. 3 (b) analyzes the distribution of the number of salient objects within the images. It is observed that images with a single salient object constitute only 8% of our dataset, whereas these with multiple salient objects (*i.e.*, the number of salient objects $\geq$ 3) constitute a significant 45.26%, surpassing the majority of existing RGB-D SOD and VSOD datasets as reported in [29, 15]. These findings highlight not only the diversity and challenge presented in our dataset but also underscore its potential for advancing saliency detection research. (3) *Video-level statistics*: we give video-level analysis on the distribution of labeled frames and the number of salient instances in the videos, with results depicted in Figs. 3 (c) and (d), respectively. These results further showcase the versatility of the proposed DViSal dataset.

4 The DVSOD Benchmark

4.1 DVSOD Baseline

To date, various network architectures [28, 83] have been developed for the tasks of RGB-D SOD and VSOD. In the former task, advanced feature fusion techniques are adopted to fuse features extracted from RGB-D images based on two-stream encoders. The latter task focuses more on exploiting temporal associations in video sequence. Based on existing techniques in RGB-D SOD [27] and VSOD [57], we assemble a straightforward DVSOD baseline, aiming at examining the advantages of utilizing depth and temporal information in improving detection accuracy & robustness.

Fig. 4 shows an overview of the proposed DVSOD baseline, which consists of four steps: 1) feature extraction; 2) fusion of RGB-D multimodal information; 3) video temporal aggregation; and 4) prediction of salient objects. Specifically, the input is a RGB-D video clip that contains a Query pair of RGB and depth images at current frame t, and L Memory pairs at past frames. We first feed these image pairs into two-stream encoders to extract RGB and depth features, respectively. Then, a multimodal fusion module, *i.e.*, cross reference modules (CRM) [27], is adopted to unify the information from two modalities at each network layer. The features across multiple layers are then decoded by the segmentation decoder [73], resulting in a series of fused multimodal features. We represent them as $\{\mathbf{f}_d \in \mathbb{R}^{H\times W\times D}\}_{d\in U}$, where $H\times W$ represents the spatial size, D is the channel dimension, and d represents the time subscript of a certain frame in the set of $U=\{t-L,\ldots,t-1,t\}$. To incorporate the temporal cues, we exploit a basic temporal fusion module, *i.e.*, the space-time memory read block [57], to furnish the Query features by engaging the rich features of Memory frames. Finally, we apply a 1×1 convolutional layer on the memory-augmented feature to predict the saliency map, which is supervised by the ground-truth saliency map using the conventional binary cross entropy (BCE) loss. More details to DVSOD baseline are described in the supplementary materials.

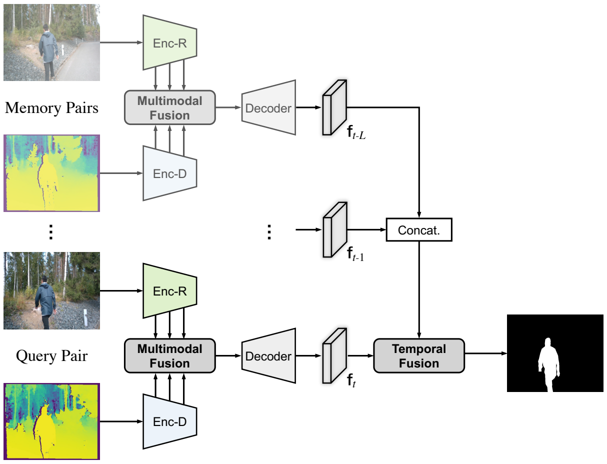


Figure 4: An overview of the DVSOD baseline.

Implementation Details. The framework is implemented with PyTorch and trained using a NVIDIA RTX A6000 GPU. We utilize the CPD decoder [73] in conjunction with the ResNet-50 encoder [21] as the base network. For the depth stream, we generate 3-channel depth maps by repeating the 1-channel depth maps. Each image is uniformly resized to 320 × 320, and we perform random horizontal flipping and cropping to avoid potential over-fitting. During model training, the learning rate is set to 1e-4, and Adam optimizer is adopted with mini-batch of 2. In the inference stage, the proposed baseline predicts saliency maps in an end-to-end manner and no post-processing procedure (*e.g.*, CRF [22]) is applied in this work.

4.2 Evaluation Metrics

We adopt four widely-used metrics to evaluate the performance of saliency models, including mean absolute error (MAE or M) [3], F-measure (F_β) [1], S-measure (S_α) [12] and E-measure (E_ξ) [13]. *The lower the MAE, the better. For other metrics, the higher score is better.* Concretely, F-measure is an overall performance measurement and is computed by the weighted harmonic mean of the precision and recall:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{1}$$

where β^2 is set to 0.3 as suggested in [1]. M represents the average absolute difference between the saliency map and ground truth. It is used to calculate how similar a normalized saliency maps $\mathcal{S} \in [0,1]^{W \times H}$ is compared to the ground truth $\mathcal{G} \in \{0,1\}^{W \times H}$:

$$M = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\mathcal{S}(x,y) - \mathcal{G}(x,y)|, \tag{2}$$

where W and H denote the width and height of $\mathcal{S}$, respectively. Structural measure (S-measure) evaluates the structural similarity between the predicted saliency maps and the binary ground truths. S-measure (denoted as S_α) contains two terms, S_o and S_r, referring to object-aware and region-aware structural similarities, respectively:

$$S_\alpha = \lambda * S_o + (1-\lambda) * S_r \tag{3}$$

where λ is the balance parameter and is set to 0.5 as in [12]. Enhanced-alignment measure (E_ξ) considers the global means of the image and local pixel matching simultaneously.

$$E_\xi = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi_s(i,j), \tag{4}$$

where $\phi_s(\cdot)$ is the enhanced alignment matrix, which reflects the correlation between $\mathcal{S}$ and $\mathcal{G}$ after subtracting their global means, respectively.

4.3 Benchmark Results

In this section, we benchmark the DVSOD task by conducting a range of experiments using the DViSal dataset with 11 popular state-of-the-art SOD methods. These methods include four ***RGB-image-based SOD models*** (PoolNet [43], BASNet [65], CPD [73], and F3Net [72]), five ***RGB-D SOD models*** (DMRA [62], CoNet [28], BBSNet [16], RD3D [7] and SPNet [95]), two *VSOD models* (PCSA [19] and UGPL [63]), and our *DVSOD baseline*. We obtained results from these methods on the new dataset by reproducing them with their publicly available codes and default setups.

Quantitative Results. In DVSOD, one important expectation compared to the RGB-based model is whether properly utilizing RGB-D and temporal features improves the per-frame segmentation accuracy. To verify it, we first conduct three comparative experiments (last three columns of Table 2) to evaluate the influence of each data type. The basic model, CPD [73], which is trained with just RGB images, achieves an overall Mean Absolute Error (MAE) score of 13.2%. When we incorporate a depth component (CPD + Dep), the error significantly decreases by 1.1%. As we employ multimodal RGB-D data and temporal video data jointly (CPD + DepVid), the performance further improves to 11.3%. These results consistently highlight the advantages of using 3D layout data and temporal contexts to locate salient objects. Considering that DVSOD is relatively new and its development is still at an early stage, we additionally present the segmentation results of related SOD/RGB-D SOD/VSOD models in Table 2, to provide observational reference for our readers.

Qualitative Results. Fig. 5 visualizes the saliency predictions from different saliency models on the DViSal dataset. We find that the model trained with both RGB-D data and video data generates saliency maps that closely match the ground truths. For example, the DVSOD baseline effectively identifies entire objects even in challenging low-light situations as displayed in Fig. 5 where other methods struggle. We attribute this success to the combined benefits of multimodal and temporal contexts. In addition, our DViSal dataset's diverse scenarios in Figs. 1-2&5 demonstrate its versatility to provide a sufficiently realistic benchmark in this field.

https://doi.org/10.52202/075280-0383

Table 2: Benchmarking results of related SOD approaches tested on the newly-proposed DViSal dataset. We report the numerical evaluation for each test subset and provide an overall measurement. $\uparrow$ & $\downarrow$ denote larger and smaller is better, respectively. The notations † and ‡ refer to RGB-D SOD and VSOD models, respectively. The best performing model is highlighted in **bold**.

*	Model	PoolNet [43]	BASNet [65]	F3Net [72]	DMRA [62]†	CoNet [28]†	BBSNet [16]†	RD3D [7]†	SPNet [95]†	PCSA [19]‡	UGPL [63]‡	CPD [73]	CPD **+Dep**†	CPD **+DepVid**†‡
DET	$E_\xi \uparrow$	.762	.761	.766	.768	.775	.784	.786	**.788**	.771	.781	.761	.781	.782
	$S_\alpha \uparrow$	.666	.672	.688	.678	.696	.692	.695	.686	.680	.688	.677	.683	**.705**
	$F_\beta \uparrow$	520	.528	.543	.532	.549	.563	.562	.565	.535	.552	.532	.554	**.568**
	$M \downarrow$	.143	.142	.146	.135	.139	.134	.127	**.126**	.134	.136	.142	.139	.129
Scene	$E_\xi \uparrow$	.769	.826	.905	.838	.830	**.906**	.852	.888	.837	.868	.804	.880	.881
	$S_\alpha \uparrow$	.646	.682	.813	.738	.772	.802	.722	.728	.738	.768	.719	.786	**.825**
	$F_\beta \uparrow$	.441	.514	.751	.571	.652	**.759**	.575	.749	.569	.721	.535	.733	.722
	$M \downarrow$	.093	.075	**.043**	.065	.062	**.043**	.065	.049	.067	.050	.078	.048	**.043**
People	$E_\xi \uparrow$	.472	.467	.517	.474	.488	.528	.515	.574	.478	.581	.409	.572	**.584**
	$S_\alpha \uparrow$	.492	.476	.510	.529	.502	.553	.551	.532	.541	.550	.463	.511	**.558**
	$F_\beta \uparrow$	.162	.114	.160	.192	.138	.178	.198	.192	.192	**.224**	.106	.169	.207
	$M \downarrow$	.153	.146	.119	.139	.128	.086	.103	.092	.128	**.078**	.254	.098	.085
Track3D	$E_\xi \uparrow$	.657	.600	.780	.634	.658	.839	.655	**.860**	.638	.838	.612	.831	.822
	$S_\alpha \uparrow$	.590	.589	**.779**	.629	.655	.724	.633	.667	.629	.693	.600	.726	.751
	$F_\beta \uparrow$	.301	.290	.567	.343	.423	.664	.369	**.682**	.342	.657	.322	.621	.609
	$M \downarrow$	.127	.145	.087	.118	.101	.080	.117	.083	.116	.080	.134	.080	**.073**
Tracklam	$E_\xi \uparrow$	.815	.798	.913	.828	.896	.923	.809	.889	.833	.924	.787	.915	**.925**
	$S_\alpha \uparrow$	.701	.682	.769	.717	.761	.793	.697	.739	.717	.785	.702	.771	**.804**
	$F_\beta \uparrow$	.622	.595	.779	.652	.751	**.809**	.628	.759	.659	.803	.608	.792	.795
	$M \downarrow$	.122	.133	.076	.114	.082	.068	.127	.084	.112	**.063**	.133	.069	**.063**
PTB	$E_\xi \uparrow$	.908	.880	.909	.893	.913	.924	.925	.911	.894	.919	.888	.906	**.927**
	$S_\alpha \uparrow$	.793	.791	.826	.810	.824	.834	.811	.793	.808	.832	.816	.819	**.842**
	$F_\beta \uparrow$	.826	.710	.795	.753	.808	.821	**.843**	.835	.754	.837	.745	.816	.833
	$M \downarrow$	.062	.067	.057	.060	.056	.054	.053	.060	.060	.052	.058	.052	**.049**
Overall	$E_\xi \uparrow$	.774	.774	.795	.783	.794	**.810**	.799	**.810**	.786	.806	.772	.805	.807
	$S_\alpha \uparrow$	.673	.679	.711	.692	.713	.715	.703	.698	.694	.709	.689	.705	**.729**
	$F_\beta \uparrow$	.537	.540	.589	.555	.585	.609	.582	.608	.557	.598	.548	.599	**.610**
	$M \downarrow$	.133	.132	.128	.125	.124	.118	.118	**.113**	.123	.119	.132	.121	**.113**

Diagnostic Analysis. Here we discuss the impact of memory size $\mathcal{M}$ and sample rate S for memory frame selection. As shown in Table 3, adding memory frames consistently improves F_β scores compared to the single-frame model (*i.e.*, $\mathcal{M} = 0$). When using more memory frames (*i.e.*, $\mathcal{M} = 3$), we see a clear performance increase (*i.e.*, 59.9%→61.0%). Increasing the $\mathcal{M}$ further beyond 3 gives marginal returns in performance. Hence, we set $\mathcal{M} = 3$ as a balance between accuracy and memory usage. We then tested different sample rates while fixing memory size $\mathcal{M} = 3$. It is discovered in Table 4 that the best performance is achieved with a moderate sample rate $S = 3$. Thus, we set both $\mathcal{M}$ and S to 3, effectively using past video frames without retaining too much outdated information.

Table 3: Ablation on the impact of memory size using F-measure and MAE.

*	$\mathcal{M}=0$	$\mathcal{M}=1$	$\mathcal{M}=2$	$\mathcal{M}=3$	$\mathcal{M}=4$
$F_\beta \uparrow$	0.599	0.603	0.607	0.610	0.608
$M \downarrow$	0.121	0.118	0.114	0.113	0.113

Table 4: Ablation on the impact of sample rate using F-measure and MAE.

*	$S=1$	$S=2$	$S=3$	$S=4$
$F_\beta \uparrow$	0.605	0.608	0.610	0.605
$M \downarrow$	0.117	0.114	0.113	0.114

5 Discussion and Outlook

In this section, we discuss several potential research problems on the emerging *DVSOD* task, and provide some feasible solutions for reference. Hopefully this could encourage more inspirations and contributions to this community. They are summarized as follows:

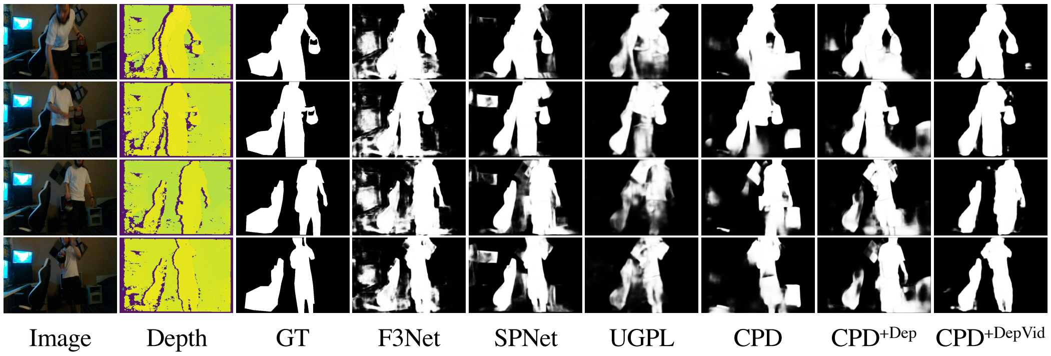


Figure 5: Visual results of different SOD models. The 'CPD$^{+\mathrm{DepVid}}$' stands for the DVSOD baseline in Sec. 4.1 that integrates both multimodal RGB-D cues and temporal contexts.

(1) ***Accuracy***: The research of DVSOD is still in its initial stage and DVSOD research has room for improvement, particularly regarding the accuracy of the DVSOD models. There's potential to enhance the accuracy of DVSOD models by incorporating ideas from the established fields of RGB, RGB-D, and Video SODs. For instance, we could integrate the multi-scale learning techniques [6, 43, 92, 88] into cross-modal and cross-frame fusion to improve the model's contextual representation. Moreover, we could introduce extra edge signals [65, 28] to aid the model in capturing object boundary details. In addition, it could be beneficial to explore more sophisticated fusion techniques [16, 95, 25] to promote effective interaction between multimodal and temporal information.
(2) ***Efficiency***: Although the engagement of RGB-D videos brings significant improvement, it introduces additional model parameters. More lightweight strategies can be explored to improve efficiency. For instance, we can develop lightweight operations such as depthwise separable convolution [10] or neural architecture search techniques [84], to advance feature extractors. On the other hand, we could explore knowledge distillation schemes [64, 54] to transfer depth structure knowledge to the RGB stream, thereby reducing the heavy overhead of the depth encoder. To evaluate the impact of lightweight strategy, we conduct an exploratory experiment by replacing the conventional convolution in our network encoder with the more efficient depth-wise separable convolution. Our experiments show a reduction in network parameters from 97.3M to 58.1M and memory usage from 17.15G to 16.21G, along with an increase in inference speed from 16.7 FPS to 21.5 FPS. However, this change leads to a drop in performance, with the F-measure falling from 0.610 to 0.575. These findings underscore the potential of efficiency-enhancing strategies, and also highlight the need to balance accuracy with efficiency in saliency model design.
(3) ***Temporal Modeling***: In this work, we conduct a preliminary investigation on the benefits of RGB-D videos for saliency detection by incorporating an well-designed temporal aggregation scheme [57] that focuses on modeling long-term dependencies. As a potential extension, we could consider introducing optical flow to capture short-term temporal cues. However, it's important to note that directly integrating optical flow could increase training complexity and potentially introduce noise, especially in complex scenes. Meanwhile, it is wroth exploring the saliency change problem (*e.g.*, the salient object changing from a person to a table) when referring to temporal information of previous frames. In the DVSOD baseline, we attempt to address it through engaging the attention mechanism within [57], which selectively discounts less relevant information and fetches the most relevant features. We encourage further investigation on this matter, and believe there is a significant opportunity for the development of more effective temporal modeling schemes that can address these challenges.
(4) ***Dataset***: As the first dataset proposed for DVSOD, there is still considerable room for improving the various aspects of the dataset such as quality, quantity, and diversity by involving a broader range of scenarios of extreme lighting conditions, as well as various types of occlusions. Furthermore, annotating such large-scale RGB-D video dataset at pixel-level is very costly in terms of both time and effort. We may explore the recent advances of photorealistic rendering techniques [11, 20] to simulate data or supplement annotations to help address challenges within the dataset.
(5) ***Evaluation Metrics***: Evaluation metrics are crucial for model training, testing, and benchmarking. However, the saliency detection community largely relies on classic metrics such as MAE and F-measure, which are not designed to assess sequential DVSOD tasks specifically. Meanwhile, the Temporal Coherence (TC) metric, frequently used in evaluating RGB video, may not accurately

https://doi.org/10.52202/075280-0383

depict DVSOD model performance due to the complex scenes within the DViSal benchmark, such as low light or cluttered backgrounds. Thus, how to design appropriate metrics specifically tailored for DVSOD remains an open challenge.

(6) ***Instance-level Extension***: As revealed in [18], instance-level extension is capable of recognizing individual instances of salient objects. It is more challenging as it requires detailed parsing within the detected salient regions, and holds significant importance for practical applications. Hence exploring the research on RGB-D video salient instance detection emerges as a promising future direction.

(7) ***Weakly-supervised Learning***: The acquirement of precise per-pixel labels is laborious and time-consuming. As an alternative, training DVSOD models with weak supervisions [80] (*e.g.*, bounding box, scribble) is an appealing research topic, which can avoid heavy annotation costs.

6 Conclusion

In this work, we present the DViSal dataset to spur research in RGB-D video salient object detection (DVSOD), a field yet to be well explored. This dataset, composed of 237 diverse RGB-D videos and comprehensive annotations, will significantly expand research avenues. Empirical benchmarking experiments demonstrate the effectiveness of multimodal video input in enhancing salient object detection. We highlight some promising future research directions and make the dataset and results publicly available to accelerate progress in this field.

Acknowledgements. This work was done jointly by UA and SRA. This research was supported by the CFI-JELF, Mitacs, University of Alberta Start-up grant, NSERC Discovery (RGPIN-2019-04575) grants.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604. IEEE, 2009.

[2] Sharon Alpert, Meirav Galun, Ronen Basri, and Achi Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *CVPR*, pages 1–8, 2007.

[3] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.

[4] Ali Borji, Dicky N Sihite, and Laurent Itti. What stands out in a scene? a study of human explicit saliency judgment. *Vision Research*, 91:62–77, 2013.

[5] Tim Caselitz, Michael Krawez, Jugesh Sundram, Mark Van Loock, and Wolfram Burgard. Camera tracking in lighting adaptable maps of indoor environments. In *ICRA*, pages 3334–3340, 2020.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018.

[7] Qian Chen, Ze Liu, Yi Zhang, Keren Fu, Qijun Zhao, and Hongwei Du. Rgb-d salient object detection via 3d convolutional neural networks. In *AAAI*, pages 1063–1071, 2021.

[8] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2014.

[9] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *Proceedings of International Conference on Internet Multimedia Computing and Service*, pages 23–27, 2014.

[10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017.

[11] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Wendelin Knauer, Klaus H Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023.

[12] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4558–4567, 2017.

[13] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pages 698–704, 2018.

https://doi.org/10.52202/075280-0383

[14] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2075–2089, 2020.

[15] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, pages 8554–8564, 2019.

[16] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *ECCV*, pages 275–292, 2020.

[17] Deng-Ping Fan, Jing Zhang, Gang Xu, Ming-Ming Cheng, and Ling Shao. Salient objects in clutter. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2344–2366, 2022.

[18] Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *CVPR*, pages 6103–6112, 2019.

[19] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*, pages 10869–10876, 2020.

[20] Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Understanding real world indoor scenes with synthetic data. In *CVPR*, pages 4077–4085, 2016.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[22] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):815–828, 2019.

[23] Wei Ji, Jingjing Li, Qi Bi, Chuan Guo, Jie Liu, and Li Cheng. Promoting saliency from depth: Deep unsupervised rgb-d saliency detection. *ICLR*, 2022.

[24] Wei Ji, Jingjing Li, Qi Bi, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05750*, 2023.

[25] Wei Ji, Jingjing Li, Cheng Bian, Zhicheng Zhang, and Li Cheng. Semanticrt: A large-scale dataset and method for robust semantic segmentation in multispectral images. In *ACM MM*, 2023.

[26] Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L Yuille, and Li Cheng. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *CVPR*, pages 1094–1104, 2023.

[27] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated rgb-d salient object detection. In *CVPR*, pages 9471–9481, 2021.

[28] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. In *ECCV*, pages 52–69, 2020.

[29] Wei Ji, Ge Yan, Jingjing Li, Yongri Piao, Shunyu Yao, Miao Zhang, Li Cheng, and Huchuan Lu. Dmra: Depth-induced multi-scale recurrent attention network for rgb-d saliency detection. *IEEE Transactions on Image Processing*, 31:2321–2336, 2022.

[30] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *CVPR*, pages 12341–12351, 2021.

[31] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *ICIP*, pages 1115–1119, 2014.

[32] Hansang Kim, Youngbae Kim, Jae-Young Sim, and Chang-Su Kim. Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE Transactions on Image Processing*, 24(8):2552–2564, 2015.

[33] Kevin Lai, Liefeng Bo, and Dieter Fox. Unsupervised feature learning for 3d scene labeling. In *ICRA*, pages 3050–3057, 2014.

[34] Minhyeok Lee, Chaewon Park, Suhwan Cho, and Sangyoun Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *ECCV*, pages 630–647, 2022.

[35] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, pages 2192–2199, 2013.

[36] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.

[37] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE Transactions on Image Processing*, 27(1):349–364, 2017.

[38] Jingjing Li, Wei Ji, Qi Bi, Cheng Yan, Miao Zhang, Yongri Piao, Huchuan Lu, et al. Joint semantic mining for weakly supervised rgb-d salient object detection. *NeurIPS*, 34:11945–11959, 2021.

[39] Jingjing Li, Wei Ji, Miao Zhang, Yongri Piao, Huchuan Lu, and Li Cheng. Delving into calibrated depth for accurate rgb-d salient object detection. *International Journal of Computer Vision*, 131(4):855–876, 2023.

[40] Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng. Exploring denoised cross-video contrast for weakly-supervised temporal action localization. In *CVPR*, pages 19914–19924, 2022.

[41] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *CVPR*, pages 2806–2813, 2014.

[42] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014.

[43] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, pages 3917–3926, 2019.

[44] Nian Liu, Ni Zhang, Ling Shao, and Junwei Han. Learning selective mutual attention and contrast for rgb-d saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9026–9042, 2021.

[45] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *ICCV*, pages 4722–4732, 2021.

[46] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *CVPR*, pages 1–8, 2007.

[47] Zhi Liu, Junhao Li, Linwei Ye, Guangling Sun, and Liquan Shen. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2527–2542, 2017.

[48] Yukang Lu, Dingyao Min, Keren Fu, and Qijun Zhao. Depth-cooperated trimodal network for video salient object detection. In *ICIP*, pages 116–120, 2022.

[49] Alan Lukezic, Ugur Kart, Jani Kapyla, Ahmed Durmush, Joni-Kristian Kamarainen, Jiri Matas, and Matej Kristan. Cdtb: A color and depth visual object tracking dataset and benchmark. In *ICCV*, pages 10013–10022, 2019.

[50] Stefan Mathe and Cristian Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *ECCV*, pages 842–856, 2012.

[51] Vida Movahedi and James H Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPRW*, pages 49–56, 2010.

[52] Munan Ning, Cheng Bian, Dong Wei, Shuang Yu, Chenglang Yuan, Yaohua Wang, Yang Guo, Kai Ma, and Yefeng Zheng. A new bidirectional unsupervised domain adaptation segmentation framework. In *IPMI*, pages 492–503, 2021.

[53] Munan Ning, Donghuan Lu, Dong Wei, Cheng Bian, Chenglang Yuan, Shuang Yu, Kai Ma, and Yefeng Zheng. Multi-anchor active domain adaptation for semantic segmentation. In *ICCV*, pages 9112–9122, 2021.

[54] Munan Ning, Donghuan Lu, Yujia Xie, Dongdong Chen, Dong Wei, Yefeng Zheng, Yonghong Tian, Shuicheng Yan, and Li Yuan. Madav2: Advanced multi-anchor based active domain adaptation segmentation. *arXiv preprint arXiv:2301.07354*, 2023.

[55] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461, 2012.

https://doi.org/10.52202/075280-0383

[56] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1187–1200, 2014.

[57] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019.

[58] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *ECCV*, pages 235–252, 2020.

[59] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020.

[60] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. RGBD salient object detection: a benchmark and algorithms. In *ECCV*, pages 92–109, 2014.

[61] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016.

[62] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7254–7263, 2019.

[63] Yongri Piao, Chenyang Lu, Miao Zhang, and Huchuan Lu. Semi-supervised video salient object detection based on uncertainty-guided pseudo labels. *NeurIPS*, 35:5614–5627, 2022.

[64] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *CVPR*, pages 9060–9069, 2020.

[65] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019.

[66] Vivek Kumar Singh, Parma Nand, Pankaj Sharma, and Preeti Kaushik. Video-based salient object detection: A survey. In *International Conference on Computing, Communication, and Intelligent Systems*, pages 617–622, 2022.

[67] Shuran Song and Jianxiong Xiao. Tracking revisited using rgbd camera: Unified benchmark and baselines. In *ICCV*, pages 233–240, 2013.

[68] Luciano Spinello and Kai O Arras. People detection in rgb-d data. In *IROS*, pages 3838–3843, 2011.

[69] Haoxiang Wang, Zhihui Li, Yang Li, Brij B Gupta, and Chang Choi. Visual saliency guided complex image retrieval. *Pattern Recognition Letters*, 130:64–72, 2020.

[70] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017.

[71] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015.

[72] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *AAAI*, pages 12321–12328, 2020.

[73] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019.

[74] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013.

[75] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, and Joni-Kristian Kämäräinen. Depth-track: Unveiling the power of rgbd tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10725–10733, 2021.

[76] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013.

[77] Jinyu Yang, Zhongqun Zhang, Zhe Li, Hyung Jin Chang, Aleš Leonardis, and Feng Zheng. Towards generic 3d tracking in rgbd videos: Benchmark and baseline. In *ECCV*, pages 112–128, 2022.

https://doi.org/10.52202/075280-0383

[78] Jianming Zhang, Shugao Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Salient object subitizing. In *CVPR*, pages 4045–4054, 2015.

[79] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. Rgb-d saliency detection via cascaded mutual information minimization. In *ICCV*, pages 4338–4347, 2021.

[80] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, pages 12546–12555, 2020.

[81] Miao Zhang, Wei Ji, Yongri Piao, Jingjing Li, Yu Zhang, Shuang Xu, and Huchuan Lu. Lfnet: Light field fusion network for salient object detection. *IEEE Transactions on Image Processing*, 29:6276–6287, 2020.

[82] Miao Zhang, Jingjing Li, Ji Wei, Yongri Piao, and Huchuan Lu. Memory-oriented decoder for light field salient object detection. *Advances in neural information processing systems*, 32, 2019.

[83] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *ICCV*, pages 1553–1563, 2021.

[84] Miao Zhang, Tingwei Liu, Yongri Piao, Shunyu Yao, and Huchuan Lu. Auto-msfnet: Search multi-scale fusion network for salient object detection. In *ACM MM*, pages 667–676, 2021.

[85] Miao Zhang, Shunyu Yao, Beiqi Hu, Yongri Piao, and Wei Ji. C^{2}dfnet: Criss-cross dynamic filter network for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 2022.

[86] Wenbo Zhang, Yao Jiang, Keren Fu, and Qijun Zhao. Bts-net: Bi-directional transfer-and-selection network for rgb-d salient object detection. In *ICME*, pages 1–6, 2021.

[87] Xiaoqi Zhao, Shijie Chang, Youwei Pang, Jiaxing Yang, Lihe Zhang, and Huchuan Lu. Adaptive multi-source predictor for zero-shot video object segmentation. *arXiv preprint arXiv:2303.10383*, 2023.

[88] Xiaoqi Zhao, Hongpeng Jia, Youwei Pang, Long Lv, Feng Tian, Lihe Zhang, Weibing Sun, and Huchuan Lu. M^{2}snet: Multi-scale in multi-scale subtraction network for medical image segmentation. *arXiv preprint arXiv:2303.10894*, 2023.

[89] Xiaoqi Zhao, Youwei Pang, Jiaxing Yang, Lihe Zhang, and Huchuan Lu. Multi-source fusion and automatic predictor selection for zero-shot video object segmentation. In *ACM MM*, pages 2645–2653, 2021.

[90] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, and Huchuan Lu. Joint learning of salient object detection, depth estimation and contour extraction. *IEEE Transactions on Image Processing*, 31:7350–7362, 2022.

[91] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pages 35–51, 2020.

[92] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Towards diverse binary segmentation via a simple yet general gated network. *arXiv preprint arXiv:2303.10396*, 2023.

[93] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic polyp segmentation via multi-scale subtraction network. In *MICCAI*, pages 120–130, 2021.

[94] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time rgb-d salient object detection. In *ECCV*, pages 646–662, 2020.

[95] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *ICCV*, pages 4681–4691, 2021.

[96] Zikun Zhou, Wenjie Pei, Xin Li, Hongpeng Wang, Feng Zheng, and Zhenyu He. Saliency-associated object tracking. In *ICCV*, pages 9866–9875, 2021.

[97] Chunbiao Zhu and Ge Li. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *ICCVW*, pages 3008–3014, 2017.

https://doi.org/10.52202/075280-0383