TENet: Triple Excitation Network for Video Salient Object Detection

Sucheng Ren[1], Chu Han[2], Xin Yang[3], Guoqiang Han[1], and Shengfeng He[⋆1][0000−0002−3802−4644]

[1] School of Computer Science and Engineering, South China University of Technology
[2] Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences
[3] Department of Computer Science and Technology, Dalian University of Technology

Abstract. In this paper, we propose a simple yet effective approach, named Triple Excitation Network, to reinforce the training of video salient object detection (VSOD) from three aspects, spatial, temporal, and online excitations. These excitation mechanisms are designed following the spirit of curriculum learning and aim to reduce learning ambiguities at the beginning of training by selectively exciting feature activations using ground truth. Then we gradually reduce the weight of ground truth excitations by a curriculum rate and replace it by a curriculum complementary map for better and faster convergence. In particular, the spatial excitation strengthens feature activations for clear object boundaries, while the temporal excitation imposes motions to emphasize spatio-temporal salient regions. Spatial and temporal excitations can combat the saliency shifting problem and conflict between spatial and temporal features of VSOD. Furthermore, our semi-curriculum learning design enables the first online refinement strategy for VSOD, which allows exciting and boosting saliency responses during testing without re-training. The proposed triple excitations can easily plug in different VSOD methods. Extensive experiments show the effectiveness of all three excitation methods and the proposed method outperforms state-of-the-art image and video salient object detection methods.

1 Introduction

When humans look into an image or a video, our visual system will unconsciously focus on the most salient region. The importance of visual saliency has been demonstrated in a bunch of applications, *e.g.*, image manipulation [32, 38], object tracking [22], person re-identification [54, 61, 62], and video understanding [44–46]. According to the slightly different goals, saliency detection can be further separated into two research interests, eye-fixation detection [20, 47] which mimics the attention mechanism of the human visual system, and salient object detection (SOD) [27, 36, 58] which focuses on segmenting the salient objects. In this paper, we focus on the latter.

Image-based salient object detection [27, 36] has been made great achievements recently. However, detecting salient objects in videos is a different story. This is because the human visual system is influenced not only by appearance but also by motion

---

⋆ Corresponding author (hesfe@scut.edu.cn).

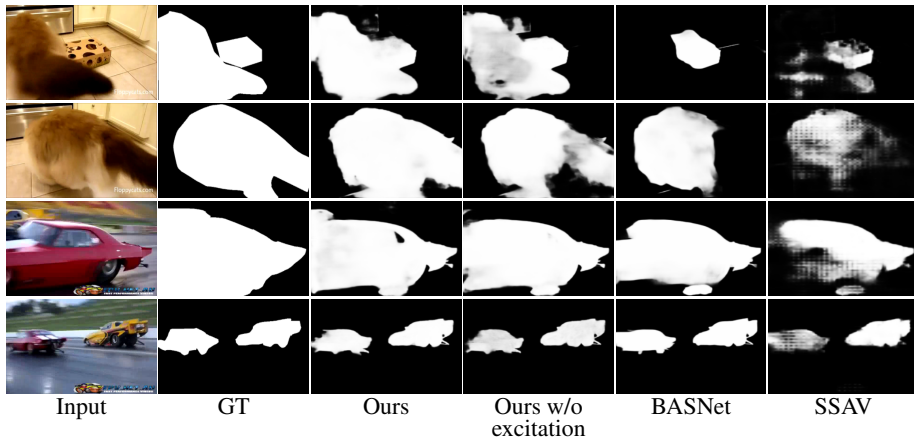| Input | GT | Ours | Ours w/o excitation | BASNet | SSAV |

Fig. 1: We propose to manually excite feature activations from three aspects, spatial and temporal excitations during training, and online excitation during testing for video salient object detection. Our simple yet effective solution injects additional spatio-temporal guidance during training and even testing for better and faster convergence. In contrast, the image-based method BASNet [36] lacks temporal understanding (first row), while the video-based method SSAV [11] suffers from spatially coarse results.

stimulus. Therefore the severe saliency shifting problem [18, 21, 41] poses the challenge in video salient object detection (VSOD). Despite different cues, *e.g.*, optical flow [11, 23, 40] and eye-fixation [11], are used to deal with this problem, the sudden shift of ground truth label makes the training difficult to converge.

Another issue in VSOD training is the contradictory features in spatial and temporal domains. As motion stimulus is a key factor of the human visual system, humans may pay attention to a moving object that does not distinct in appearance. Although features fusion is typically applied, extracting temporal features is much more difficult than spatial ones, as motion blurring, object, and camera movements are involved. They cannot capture clear object boundaries as the spatial features do. As a consequence, VSOD methods (*e.g.*, the last column of Fig. 1) produce spatially coarse results in the scenarios with moving objects. We argue that a simple feature fusion strategy cannot solve this problem, and alternative guidance during training is desired.

To address the above two issues, we propose a Triple Excitation Network (TENet) for video salient object detection. Three types of excitations are tailored for VSOD, *i.e.*, spatial and temporal excitations during training, and online excitation during testing. We adopt a similar spirit with curriculum learning [2], that we aim to loosen the training difficulties at the beginning by exciting selective feature activations using ground truth, then gradually increase task difficulty by replacing such ground truth by our learnable complementary maps. This strategy simplifies the training process of VSOD and boosts the performance with faster convergence and we name it as semi-curriculum learning. In particular, spatial excitation learns spatial features to obtain a boundary-sharp segment. While the temporal excitation aims to leverage previous predictions and excites spatio-temporal salient regions from a spatial excitation map and an optical flow. These excitations are directly performed on the activations of features, and therefore provide

direct supports on mitigating errors brought by the problems of saliency shifting and inaccurate temporal features. Thanks to our semi-curriculum learning design, we can apply excitations in testing by proposing online excitation which can be done without any further training. It is worth noting that the proposed excitation mechanism can easily plug in different VSOD methods. Extensive experiments are performed to qualitatively and quantitatively evaluate the effectiveness of the proposed method, it outperforms state-of-the-art methods on four VSOD benchmarks.

The main contributions of this paper are four-fold:

– We delve into the problems of saliency shifting and inaccurate temporal features and tailor a triple excitation mechanism to excite spatio-temporal features for mitigating the training difficulties caused by these two problems. Better and faster convergence can be obtained.
– We present a semi-curriculum learning strategy for VSOD. It reduces the learning ambiguities by exciting certain activations during the beginning of training, then gradually reduces the curriculum rate to zero and transfers the weight of excitation from ground truth to our learnable complementary maps. This learning strategy progressively weans the network off dependence on ground truth, which is not only beneficial for training but also for testing.
– We propose an online excitation that allows the network to keep self-refining during the testing phase.
– We outperform state-of-the-art SOD and VSOD methods on four benchmarks.

2   Related Works

Image Salient Object Detection. Traditional image saliency detection methods [6, 37, 43] usually rely on the hand-crafted features, *e.g.*, color contrast, brightness. It can be separated into two categories, bottom-up [3, 13, 19] and top-down [3, 53] approaches. Driven by a large amount of labeled data, researchers attempt to consider saliency detection as a classification problem [24, 60] by simply classifying the patches into non-salient or salient. However, the patches cropped from the original image are usually small and lack of global information. Recent approaches adopt FCN [30] as a basic architecture to detect saliency in an end-to-end manner. Based on that, edge information is incorporated to promote clear object boundaries, by a boundary-enhanced loss [12] or jointly trained with edge detection [27]. Attention mechanism [28, 63] is also introduced to filter out a cluttered background. All these methods provide a useful guideline to handle spatial information.

Video Salient Object Detection. The involved temporal information makes video salient object detection much harder than image salient object detection. Some existing approaches try to fuse spatial and temporal information using graph cut [26, 33], gradient flow [48], and low-rank coherence [6]. Researchers also try to associate spatial with temporal information using deep networks. Wang *et al*. [49] concatenate the current frame and saliency of the previous frame to process temporal information. Li *et al*. [23] propose to use optical flow to guide the recurrent neural encoder. They use ConvLSTM to process optical flow and warp latent features before feeding into another ConvLSTM. To capture a wider range of temporal information, a deeper bi-directional

ConvLSTM [40] has been proposed. Fan *et al.* [11] mitigate the saliency shifting problem by introducing the eye-fixation mechanism. Similar to VSOD, unsupervised video object segmentation [31, 50, 55] aims at segmenting primarily objects with temporal information. However, as discussed above, replying only to additional features cannot solve problems of saliency shifting and inaccurate temporal features well. We resolve them from the perspective of reducing training difficulties.

Extra Guidance for CNNs. Introducing extra guidance is a popular solution to aid the training of a deep network. For example, jointly training semantic segmentation and object detection [5, 8, 14] improves the performances for both tasks. One limitation of multi-task training is that it needs two types of annotations. Some other works introduce two different types of annotations from the same task, *e.g.*, box and segmentation labels for semantic segmentation [15, 59], to boost the training performance. Different from these methods, we do not introduce extra task or annotation for training, but directly employ ground truth of the same task as well as pseudo-label for exciting features activations.

3   Triple Excitation Network

3.1   System Overview

Given a series of frames $\{T_n | n = 1, 2, ..., N\}$, we aim to predict the salient object in frame T_n. Fig. 2 shows the pipeline of our proposed Triple Excitation Network. The basic network is an encoder-decoder architecture with skip connections (which are hidden in Fig. 2 for simplification). Our framework consists of three branches with respective purposes. The spatial excitation prediction branch is proposed to predict spatial complementary maps with rich spatial information for generating spatial excitation map. The temporal excitation prediction branch leverages the optical flow and spatial excitation maps to generate the temporal complementary maps. These two complementary maps are combined with ground truth to provide the additional guidance for the network training and testing. ConvLSTM [39] is introduced to inject temporal information on the feature maps extracted from the encoder in video saliency prediction branch. After the spatial and temporal excitations, the final saliency map of the current frame T_n is generated by the saliency decoder. During the testing phase, the final saliency map is further adopted for online excitation.

3.2   Excitation Module

Due to the difficulties of handling saliency shifting and the contradictory features in spatial and temporal domains, a simple feature fusion strategy is no longer sufficient for video salient object detection. Therefore, we propose an excitation mechanism shown in Fig. 3 as the additional guidance to reinforce certain feature activations during training.

It is worth noting that our proposed excitation mechanism is a super lightweight plug-in that does not waste computational power because it does not involve any convolution operation. Given an input tensor M, an excitation map E with the pixel values
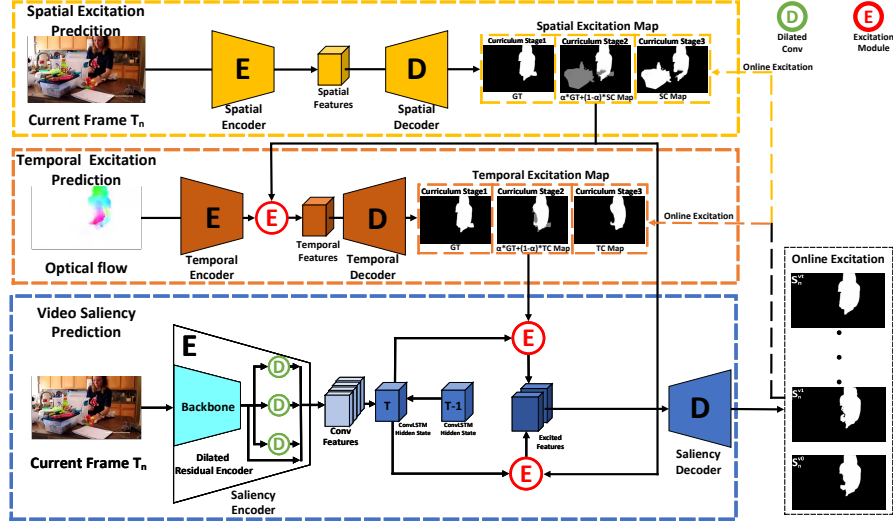
Fig. 2: Network architecture of TENet. The upper two branches provide spatial and temporal excitation with a curriculum learning strategy. In each curriculum stage, we balance the contribution of the ground truth and the complementary map to avoid the overdependency of ground truth during training phase. ConvLSTM is applied in the third branch to introduce the temporal features from the previous frames. During testing phase, an optional online excitation allows the network to keep refining the saliency prediction results by keeping updating the complementary maps with previous predictions recurrently. Note that, the structures of all the encoders (E) in the network are the exactly the same but with different parameters. We only show the saliency encoder for simplification.

under the range $[0, 1]$, we can obtain the output excited tensor M' by the following equation:

$$M' = \beta \odot E \odot M + (1 - \beta) \odot M, \tag{1}$$

where $\odot$ is element-wise multiplication. β is a learnable excitation rate which determines the intensity of excitation based on the feedback of the model itself. It learns an optimum balance between manual excitation and learned activations.

Semi-curriculum Learning. The excitation map can actually be any map that reflects the feature responses required excitation. It can be the ground truth saliency map in this application. However, directly utilizing the ground truth for excitation may let the network excessively rely on the ground truth itself. Therefore, we introduce a semi-curriculum learning strategy for our excitation mechanism. This strategy shares a similar spirit to the curriculum learning framework [2], in which they argue that training with an easy task at first then continues with more difficult tasks gradually may leads to better optimization. Therefore, as the training goes on, we update the excitation map by trading off the intensity between the ground truth GT and a learnable complementary
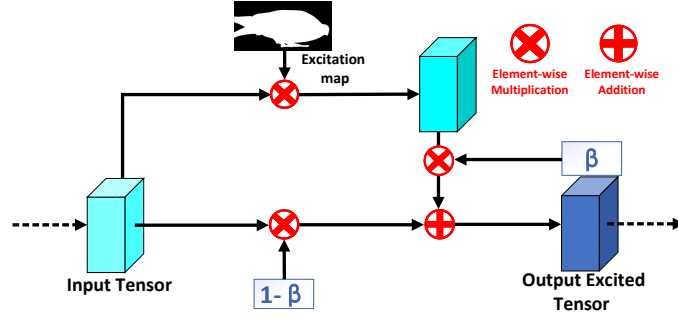
Fig. 3: The proposed excitation module. It strengthens saliency features responses by manually exciting certain feature activations, and the training is controlled by a learnable excitation rate β automatically adjusted according to the feedback of neural network.

map S as follows:

$$E = \alpha \odot GT + (1 - \alpha) \odot \mathcal{S} \tag{2}$$

where α is the curriculum rate which has been initially set as 1 and will automatically decay to 0 to transfer the contribution from the ground truth to the learnable complementary map. The gradually decreased curriculum rate α will increase the task difficulty and thus can effectively avoid the overdependence on the ground truth. In the meanwhile, this is also the key to enable online excitation.

In practice, we divide the training process into three curriculum stages along with the change of training epoch e. The excitation map E in Eq. 2 can be reformulated as follows:

$$E = \begin{cases} GT & \text{Stage1} : e \leq 2 \\ \alpha \odot GT + (1 - \alpha) \odot \mathcal{S} & \text{Stage2} : 2 < e \leq 10 \\ \mathcal{S} & \text{Stage3} : e > 10 \end{cases} \tag{3}$$

Stage 1: Due to the imbalance foreground and background pixels in VSOD, for example, the salient pixels take only 8.089% in the DAVIS Dataset [35], reinforcing the network to focus on salient region by using the ground truth as the excitation map provides a shortcut for a better optimization at the beginning of training.

Stage 2: However, models tend to rely on the perfect ground truth and it degrades the model performance once the ground truth is removed. As a result, we gradually replace the ground truth by our learned complementary map (controlled by the curriculum rate α). During this period, the predicted complementary map is to inject perturbation and prevent the model from too sensitive to the perfect ground truth.

Stage 3: When α decays to zero, our model is excited only by the complementary maps. This avoids our network overdependence on GT, more importantly, this is the key to enabling online excitation. We keep training the network in this stage to 15 epochs.

3.3 Spatial-temporal Excited Saliency Prediction

Our model consists of three branches, the first two are for generating excitation maps, and we predict the video saliency result in the third branch by predicting the video

frames one by one. For all these branches, we extract the feature maps by a dilated residual encoder as described below.

Dilated Residual Encoder. The backbone of the encoder borrows from ResNet [16]. We replace the first convolutional layer and the following pooling layer by a 3×3 convolutional layer with stride 1 and extract the deep features $\mathcal{X}_n \in \mathbb{R}^{w \times h \times c}$. To handle the uncertain scales of the objects, we extract the multi-level feature maps by introducing dilated convolution and keep increasing the dilation rates $\{2^k\}_{k=1}^K$. The k-th level features extracted from the dilated convolution with dilation rate 2^k is $\mathcal{D}^k \in \mathbb{R}^{w \times h \times c'}$. The output feature maps $\mathcal{F}$ is the concatenation of all the outputs from dilated residual encoder:

$$\mathcal{F}_n = \left[\mathcal{X}_n, \mathcal{D}_n^1, \mathcal{D}_n^2, ..., \mathcal{D}_n^K \right], \tag{4}$$

where $\mathcal{F} \in \mathbb{R}^{w \times h \times (c+K*c')}$. The feature maps $\mathcal{F}_n$ from dilated residual encoder not only keeps the original features $\mathcal{X}_n$ but also covers much larger receptive fields with local-global information. All the encoders in the three branches share the same structure but have different parameters.

Spatial Excitation Prediction Branch. We predict the spatial excitation map from a single frame in the spatial excitation branch which has an encoder-decoder structure. We use the dilated residual encoder as mentioned above, and the decoder has four convolutional stages. Each stage contains three convolutional blocks, each of which is a combination of a convolutional layer, a batch normalization layer, and a ReLU activation layer. With the spatial complementary map S_n^s generated by the spatial decoder, we can calculate the spatial excitation map E_n^s by Eq. 3.

Temporal Excitation Prediction Branch. The temporal branch is designed to tackle the human visual attention shifting problem. It takes an input optical flow, which is calculated by the state-of-the-art optical flow prediction method [29], and outputs a temporal excitation map. This branch has the same network structure with the spatial one. The difference is, we make a spatial excitation on the latent features in the temporal branch in order to associate the temporal excitation with the spatial excitation.

Given the input optical flow from the frame T_{n-1} to the frame T_n, we have the latent features $\mathcal{F}_n^t$ extracted from the temporal encoder. Then the spatial excitation maps of the two consecutive frames E_n^s and E_{n-1}^s are combined together for spatial excitation. The excited temporal latent features $\mathcal{F'}_n^t$ is calculated as follows:

$$\mathcal{F'}_n^t = \beta^{t \to s} \odot clip(E_n^s + E_{n-1}^s) \odot \mathcal{F}_n^t + (1 - \beta^{t \to s}) \odot \mathcal{F}_n^t, \tag{5}$$

where $clip(\cdot)$ operation clips the value between 0 and 1 and β^t is a learnable temporal excitation rate. The optical flow reveals the moving objects explicitly and the predicted temporal complementary map covers temporally salient regions for governing training. With the temporal complementary map $\mathcal{S}_n^t$ generated by the temporal decoder, we can calculate the temporal excitation map E_n^t by Eq. 3.

Video Saliency Prediction Branch. In this branch, we aim to predict saliency maps with spatially sharpen the object boundaries by leveraging the spatio-temporal excitation mechanism. After the feature extraction with dilated residual encoder, We apply the bi-directional ConvLSTM on the extracted feature maps $\mathcal{F}_n^v$ to obtain long-short term spatial and temporal features:

$$\overrightarrow{\mathcal{H}}_n^v = ConvLSTM(\mathcal{F}_n^v, \overrightarrow{\mathcal{H}}_{n-1}^v), \tag{6}$$

$$\overleftarrow{\mathcal{H}}_n^v = ConvLSTM(\mathcal{F}_n^v, \overleftarrow{\mathcal{H}}_{n-1}^v). \tag{7}$$

We consolidate the features representations on two temporal directions by leveraging both spatial excitation and temporal excitation and obtain bi-directional feature maps $\overrightarrow{\mathcal{H}'}_n^v$ and $\overleftarrow{\mathcal{H}'}_n^v$ of frame T_n as follows:

$$\begin{aligned}\overrightarrow{\mathcal{H}'}_n^v = cat(&\beta^{s \to v} \odot E_n^s \odot \overrightarrow{\mathcal{H}}_n^v + (1 - \beta^{s \to v}) \odot \overrightarrow{\mathcal{H}}_n^v \\ &+ \beta^{t \to v} \odot E_n^t \odot \overrightarrow{\mathcal{H}}_n^v + (1 - \beta^{t \to v}) \odot \overrightarrow{\mathcal{H}}_n^v),\end{aligned} \tag{8}$$

$$\begin{aligned}\overleftarrow{\mathcal{H}'}_n^v = cat(&\beta^{s \to v} \odot E_n^s \odot \overleftarrow{\mathcal{H}}_n^v + (1 - \beta^{s \to v}) \odot \overleftarrow{\mathcal{H}}_n^v \\ &+ \beta^{t \to v} \odot E_n^t \odot \overleftarrow{\mathcal{H}}_n^v + (1 - \beta^{t \to v}) \odot \overleftarrow{\mathcal{H}'}_n^v),\end{aligned} \tag{9}$$

where E_n^s and E_n^t are the spatial and temporal excitation maps respectively and $cat(\cdot, \cdot)$ is the concatenation operation. $\beta^{\cdot}$ are the learnable parameters. Since we perform the excitation strategy on the latent space, the excitation maps will be first downsampled to the same size with the feature maps.

The bi-directional excited hidden stage is then concatenated together to produce the final saliency result $\mathcal{S}_n^v$ of the frame T_n by the saliency decoder W^s:

$$\mathcal{S}_n^v = W^s \otimes cat(\overrightarrow{\mathcal{H}'}_n^v, \overleftarrow{\mathcal{H}'}_n^v). \tag{10}$$

Note that the saliency decoder W^s here has the same structure but different parameters with temporal and spatial decoders.

3.4   Loss Function

We borrow the loss function from BASNet [36]. It includes the cross entropy loss [9], SSIM loss [51], and IoU loss [56]. They measure the quality of saliency map in pixel-level, patch-level, and object-level respectively.

$$l = l_{bce} + l_{ssim} + l_{IoU}. \tag{11}$$

The cross entropy loss l_{bce} measures the distance between two probability distributions which is the most common loss function in binary classification and salient object detection.

$$l_{bce}(\mathcal{S}, GT) = -\sum_{i=1}^{w} \sum_{j=1}^{h} [GT(i,j) \log \mathcal{S}(i,j) + (1 - GT(i,j)) \log(1 - \mathcal{S}(i,j))], \tag{12}$$

where S is the network predicted saliency map, S_{gt} is the ground truth saliency map.

The SSIM is originally designed for measuring the structural similarity of two images. When applying this loss into saliency detection, it helps the network pay more attention to the object boundary due to the higher SSIM activation around the boundary. Let $x = \left\{x_i : i = 1, \cdots, N^2\right\}, y = \left\{y_i : i = 1, \cdots, N^2\right\}$ be the corresponding N×N patches of predict saliency and ground truth label respectively, we have:

$$l_{ssim}(\mathcal{S}, GT) = 1 - \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{13}$$

where μ_x, μ_y and σ_x, σ_y are the mean and variance, and σ_{xy} is the co-variance. $c_1 = 0.01^2$, $c_2 = 0.03^2$ are constants for maintaining stability.

Intersection over Union (IoU) is widely used in detection and segmentation for evaluation and also used as training loss. The IoU loss is defined as:

$$l_{iou}(\mathcal{S}, GT) = 1 - \frac{\sum\limits_{i=1}^{w}\sum\limits_{j=1}^{h} \mathcal{S}(i,j)GT(i,j)}{\sum\limits_{i=1}^{w}\sum\limits_{j=1}^{h} [\mathcal{S}(i,j) + GT(i,j) - \mathcal{S}(i,j)GT(i,j)]}. \tag{14}$$

The above losses apply to three branches, and the total objective function of our network is the combination of the spatial excitation loss $l(\mathcal{S}^s, GT)$, temporal excitation loss $l(\mathcal{S}^t, GT)$, and the video saliency loss $l(\mathcal{S}^v, GT)$:

$$\mathcal{L} = l(\mathcal{S}^s, GT) + l(\mathcal{S}^t, GT) + l(\mathcal{S}^v, GT). \tag{15}$$

3.5 Online Excitation

In our network design, the quality of the excitation map plays an important role in final saliency map prediction. During training, we use a predicted excitation map for highlighting salient activations in the features. Thanks to our semi-curriculum learning strategy, the excitation map does not rely on GT during the testing phase, and we can use a better excitation map to replace the initial guidance. In this way, we design an additional excitation strategy in the testing phase to refine the predicted saliency map without any further training, we call it online excitation. Users can refine the saliency result by recurrently replacing the excitation maps with previous video saliency prediction outputs for better guidance. It provides an additional option for the users to trade-off between the saliency prediction quality and the computational cost during testing. Theoretically, if the excitation map is the same as the ground truth saliency map, our network can give the optimal solution of saliency prediction. We have conducted an experiment in the ablation study to prove the effectiveness of our online excitation.

4 Experiments

4.1 Implementation Details

Our method is trained with three datasets (following the strategy of Fan *et al.* [11]). Images are loaded into a batch according to their dataset, and we alternately train the

Spatial Excitation branch with images from DUT-OMRON [52] and DAVIS [35], Temporal Excitation branch with optical flow from DAVIS and DAVSOD [11], and the whole model with video from DAVIS and DAVSOD. The optimizer is SGD with momentum 0.9 and weight decay 0.0005. The learning rate starts from 5e-4 and decay to 1e-6. The curriculum rate initially set to 1, and decays following a cosine function. All inputs are randomly flip horizontally and vertically for data argumentation. During training, every input image will be resized to 256×256 and randomly center cropped to 224×224. All the resizing method is the bilinear interpolation. It takes about 40 hours to converge, which is one and a half times shorter (80 hours) than training without excitation. It shows that our excitation not only boosts the performance as shown below but also accelerates the training process.

4.2   Datasets

We conduct the experiments on four most frequently used VSOD datasets, including Freiburg-Berkeley motion segmentation dataset (FBMS) [4], video salient object detection dataset (ViSal) [48], densely annotation video segmentation dataset (DAVIS) [35], and densely annotation video salient object detection dataset (DAVSOD) [11]. **FBMS** contains 59 videos with only 720 annotated frames. There are 29 videos for training and the rest of them are for testing. **DAVIS** is a high quality and high resolution densely annotated dataset under two resolutions, 480p and 1080p. There are 50 video sequences with 3455 densely annotated frames in pixel level. 30 videos with 2079 frames are for training and 20 videos with 1376 frames are for validation. **ViSal** is the first dataset specially designed for video salient object detection which includes 17 videos and 193 manual annotated frames. **DAVSOD** is the latest and most challenging video salient detection dataset with pixel-wise annotations and eye-fixation labels. We follow the same setting of SSAV [11] and evaluate on 35 test videos.

4.3   Evaluation Metrics

We take three measurements to evaluate our methods: MAE [34], F-measure [1], Structural measurement (S-measure) [10]. MAE is the mean absolute value between predicted saliency map and the groundtruth:

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{w} \sum_{j=1}^{h} |M(i,j) - G_s(i,j)|.$$ (16)

In testing, MAE value is the average MAE over the whole testing set.

F_β takes both precision and recall into consideration:

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall},$$ (17)

where β^2 is usually set to 0.3 and we use maximum F_β for evaluation.

S-measure takes both region and object structural similarity into consideration:

$$S = \mu * S_o + (1 - \mu) * S_r,$$ (18)

Method		FBMS			ViSal			DAVIS			DAVSOD		
		MAE↓	max F_β ↑	S ↑	MAE↓	max F_β ↑	S ↑	MAE↓	max F_β ↑	S ↑	MAE↓	max F_β ↑	S ↑
DSS [17]	I	0.080	0.760	0.831	0.024	0.917	0.925	0.059	0.720	0.791	0.112	0.545	0.630
BMPM [57]	I	0.056	0.791	0.844	0.022	0.925	0.930	0.046	0.769	0.834	0.089	0.599	0.704
BASNet [36]	I	0.051	0.817	0.861	**0.011**	**0.949**	0.945	0.029	0.818	0.862	0.110	0.597	0.670
SIVM [37]	V	0.233	0.416	0.551	0.199	0.521	0.611	0.211	0.461	0.551	0.291	0.299	0.491
MSTM [43]	V	0.177	0.501	0.617	0.091	0.681	0.744	0.166	0.437	0.588	0.210	0.341	0.529
SFLR [6]	V	0.119	0.665	0.690	0.059	0.782	0.815	0.055	0.726	0.781	0.132	0.477	0.627
SCOM [7]	V	0.078	0.796	0.789	0.110	0.829	0.761	0.048	0.789	0.836	0.217	0.461	0.603
SCNN [42]	V	0.091	0.766	0.799	0.072	0.833	0.850	0.066	0.711	0.785	0.129	0.533	0.677
FCNS [49]	V	0.095	0.745	0.788	0.045	0.851	0.879	0.055	0.711	0.781	0.121	0.545	0.664
FGRNE [23]	V	0.085	0.771	0.811	0.041	0.850	0.861	0.043	0.782	0.840	0.099	0.577	0.701
PDBM [40]	V	0.066	0.801	0.845	0.022	0.916	0.929	0.028	0.850	0.880	0.107	0.585	0.699
SSAV [11]	V	0.044	0.855	0.873	0.018	0.939	0.943	0.029	0.861	0.891	0.092	0.602	0.719
MGAN [25]	V	0.028	0.889	0.907	0.015	0.944	0.944	0.022	0.897	0.911	0.080	0.637	0.740
Ours	V	0.027	0.887	0.910	0.014	0.947	0.943	0.021	0.894	0.905	0.078	0.648	0.753
Ours★	V	0.026	0.897	0.915	0.014	0.949	0.946	0.019	0.904	0.916	0.074	0.664	0.780

Table 1: Quantitative comparison with image salient object detection methods (labeled as I) and the state-of-the-art VSOD methods (labelled as V) by three evaluation metrics. $Ours$ and $Ours^\star$ indicate the results without and with the online excitation. Top three performances are marked in Red, Green, and Blue respectively.

where S_0 and S_r denotes the region-aware structural similarity and object-aware structural similarity respectively. μ is set to 0.5.

4.4   Comparisons with State-of-the-arts

We compare our method with 13 saliency methods, including image salient object detection methods (DSS [17], BMPM [57], BASNet [36]) and video salient object detection methods (SIVM [37], MSTM [43], SFLR [6], SCOM [7], SCNN [42], FCNS [49], FGRNE [23], PDBM [40], SSAV [11], MGAN [25]).

Table 1 show the quantitative evaluation with existing methods. For image salient object detection methods, even without the help of temporal information, they perform well in some video salient datasets. That is because objects that distinct in appearance draws the most attention from the viewer if they are not moving dramatically in the video. However, these methods are not comparable with video-based methods due to the lack of temporal consideration. On the other hand, although the most recent VSOD method SSAV [11] leverages eye-fixation information to guide the network, the imbalance of spatial and temporal domains harm the accuracy of their saliency results. Our method performs the best statistics results among all the methods in all datasets. Note that, we have shown two results produced by our network, '$Ours$' and '$Ours^\star$'. '$Ours$' is the model with the excitation map generated by the excitation prediction branch, *i.e.*, without online excitation. '$Ours^\star$' indicates the results involve online excitation by recurrently applying the previous network outputs. We can find a significant improvement when we apply online excitation. It proves that a precise excitation map can give more accurate guidance for the network, even without further training.

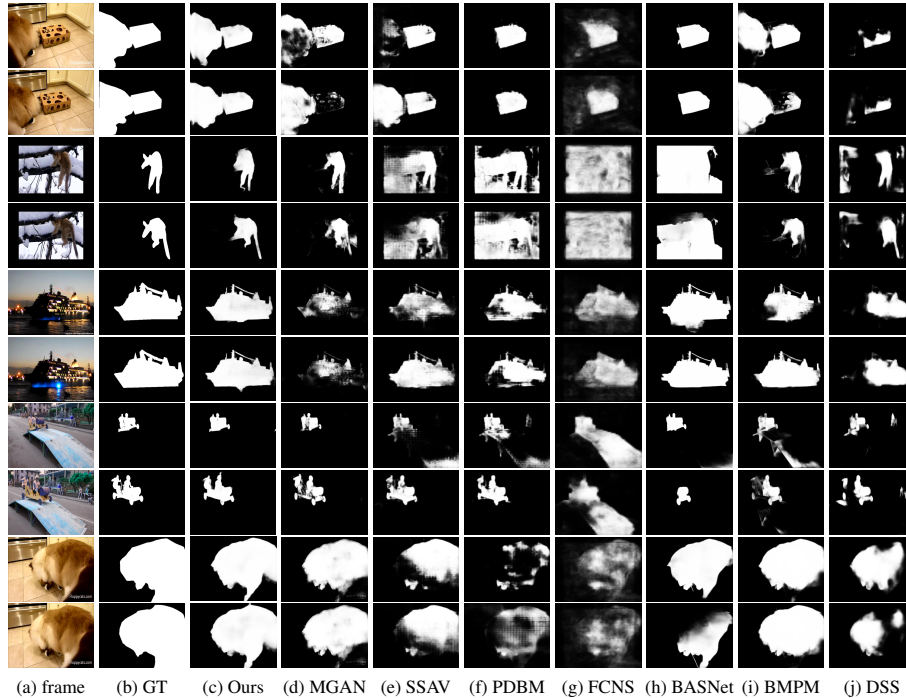| (a) frame | (b) GT | (c) Ours | (d) MGAN | (e) SSAV | (f) PDBM | (g) FCNS | (h) BASNet | (i) BMPM | (j) DSS |

Fig. 4: Qualitative comparison with state-of-the-art methods. Our TENet produces clear object boundaries while capture temporally salient objects in the video.

Another interesting observation that can be found in Table 1 is that the datasets also affect the network performance a lot. Because when looking into a video, people tend to focus on moving objects. Some moving objects are not salient in the single frame will definitely distract the network. For some easy dataset whose salient objects are moving and occupy a large part of the image, like *ViSal*, the performance is good for both image-based methods and video-based methods. While in some complicated datasets like *DAVSOD* and *FBMS*, the salient objects in the temporal domain are unfortunately not salient in the spatial domain. The statistical results of them are much worse than those easier datasets. Since our proposed excitation mechanism governs both the spatial and temporal information, our methods gain a much higher performance than the existing methods.

Fig. 4 shows the qualitative comparison. The results predicted by image-based methods shown in Fig. 4 (h)-(j) fails to detect the object region accurately and sometimes cannot distinguish the foreground and background due to the lack of temporal information. VSOD methods shown in Fig. 4 (d)-(g) provide visually more reasonable saliency maps. However, the boundary of the salient object is not clear and the inside region is blurry, due to the contradictory spatial and temporal features. In contrast, our results (without online excitation, see Fig. 4(c)) show clear boundaries as well as high-confidence interior salient regions. Our model can produce the closest saliency map to the ground truth.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
Temporal	✓	✓	✓	✓					✓	✓	✓	✓
Spatial					✓	✓	✓	✓	✓	✓	✓	✓
Online (1)		✓				✓				✓		
Online (20)			✓				✓				✓	
Online (GT)				✓				✓				✓
Results — MAE	0.092	0.090	0.091	0.069	0.084	0.081	0.080	0.062	0.078	0.075	0.074	0.053
Results — F_β	0.591	0.595	0.594	0.691	0.615	0.628	0.631	0.688	0.648	0.659	0.664	0.841
Results — S	0.693	0.702	0.708	0.738	0.715	0.733	0.741	0.764	0.753	0.772	0.780	0.862

Table 2: Ablation study for triple excitation mechanism on the DAVSOD dataset. We separately demonstrate the effectiveness of each excitation component. Online (N) indicates the online excitation with N iterations.

4.5   Ablation Study

In this section, we explore the effectiveness of our proposed modules. We test the performance on DAVSOD which is the most challenging VSOD dataset.

Effectiveness of Triple Excitation. Table 2 shows an ablation study to evaluate the effectiveness of our triple excitation method. In this experiment, we choose 14 configurations of different excitation strategies. The checkmark in Table 2 indicates the activated excitation component. We can observe that both temporal and spatial excitations boost the detection performance by a large margin (comparing to the first column without checkmark).

We then demonstrate the proposed online excitation. Ideally, we can keep refining the excitation map until convergence. We perform online excitation for one and multiple iterations, labeled as $Online(1)$ and $Online(20)$. In our experiment, applying more than 20 iterations cannot bring extra improvement. We also show the ideal case, in which uses the ground truth saliency map as the excitation map, labeled as $Online(GT)$ in Table 2. Although using ground truth as excitation information is impossible in testing, we demonstrate the upper-bound of our method.

The statistical results reveal that our three excitations all together work well as we expected. In our original design, online excitation plays an important role in our proposed system. It not only introduces more precise excitation guidance but also provides additional optional to exchange the computational cost for the prediction accuracy, by iteratively running the online excitation. $Online(20)$ shows the results of 20 times online iterations. We can find an obvious improvement in three measurements. '$Ours^\star$' in Table 1 is implemented with 20 iterations. Furthermore, the performance is significantly boosted when we feed the ground truth saliency map into online excitation. It proves that a more accuracy excitation map will bring more profits which also implicitly demonstrates the effectiveness of the excitation mechanism.

Effectiveness of Semi-curriculum Learning. In here we verify the effectiveness of our semi-curriculum learning. Our semi-curriculum learning involves two main com-

Method	DAVSOD			
	MAE↓	max F_β ↑	S ↑	Convergence Time
Baseline	0.112	0.579	0.694	80 hours
Baseline + Curriculum	0.108	0.584	0.699	32 hours
Baseline + Learned Excitation	0.080	0.641	0.743	46 (pre-training) + 38 hours
Baseline + Semi-curriculum	0.078	0.648	0.753	40 hours

Table 3: Ablation study on the proposed semi-curriculum learning.

Method	SIVM [37]	BMPM [57]	FCNS [49]	PMDB [40]	SSAV [11]	Ours
Time(s)	18.1	0.03	0.50	0.08	0.08	0.06

Table 4: Running time comparison of existing methods.

ponents, GT and learned complementary maps. We consider using the GT only as the traditional curriculum solution, and we also compare to only using the learned complementary maps for excitation. As can be seen in Table 3, the traditional curriculum learning hugely reduces the convergence time. However, as the network relies too much on the perfect ground truth, once no guidance is provided during testing, the curriculum learning strategy does not bring too much improvement. On the other hand, using a learned complementary map for excitation can ease this problem. To provide initial supervision of the network, we pretrain the complementary map for static saliency detection. Using the learned complementary map can provide guidance during testing, which is the key to maintain consistent performances between training and testing phrases. The main limitation is that it requires a separated pretraining, which largely increases the convergence time. The proposed semi-curriculum learning strategy remedies the limitations of previous two methods, leading to faster and better convergence.

4.6   Timing Statistics

We also show the running time of different models in Table 4. All the methods are tested on the same platform: Intel(R) Xeon(R) CPU E5-2620v4 @2.10GHz and GTX1080Ti. The timing statistics do not include the pre-/post-processing time. Due to our plug-and-play excitations, we can have a fast timing performance compared with most of the deep learning based VSOD methods.

5   Conclusion

This paper proposes a novel video salient object detection method equipped with a triple excitation mechanism. Spatial and temporal excitations are proposed during training phase to tackle the saliency shifting and contradictory spatio-temporal features problems. Besides, we introduce semi-curriculum learning during training to loosen the task difficulty at first and reach a better converage. Furthermore, we propose the first online excitation in testing phase to allow the network keep refining the saliency result by using the network output saliency map for excitation. Extensive experiments show that our results outperform all the competitors.

Acknowledgement

References

1. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: CVPR. pp. 1597–1604 (2009) 10
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML. pp. 41–48 (2009) 2, 5
3. Borji, A.: Boosting bottom-up and top-down visual features for saliency estimation. In: CVPR. pp. 438–445. IEEE (2012) 3
4. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: ECCV. pp. 282–295 (2010) 10
5. Cao, J., Pang, Y., Li, X.: Triply supervised decoder networks for joint detection and segmentation. In: CVPR. pp. 7392–7401 (2019) 4
6. Chen, C., Li, S., Wang, Y., Qin, H., Hao, A.: Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. IEEE TIP **26**(7), 3156–3170 (2017) 3, 11
7. Chen, Y., Zou, W., Tang, Y., Li, X., Xu, C., Komodakis, N.: Scom: Spatiotemporal constrained optimization for salient object detection. IEEE TIP **27**(7), 3345–3357 (2018) 11
8. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: CVPR. pp. 3150–3158 (2016) 4
9. De Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Annals of operations research **134**(1), 19–67 (2005) 8
10. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV. pp. 4548–4557 (2017) 10
11. Fan, D.P., Wang, W., Cheng, M.M., Shen, J.: Shifting more attention to video salient object detection. In: CVPR. pp. 8554–8564 (2019) 2, 4, 9, 10, 11, 14
12. Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: CVPR (2019) 3
13. Gao, D., Vasconcelos, N.: Bottom-up saliency is a discriminant process. In: ICCV. pp. 1–6 (2007) 3
14. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous detection and segmentation. In: ECCV. pp. 297–312 (2014) 4
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2961–2969 (2017) 4
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 7
17. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. In: CVPR. pp. 3203–3212 (2017) 11
18. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE TPAMI (11), 1254–1259 (1998) 2

16   S. Ren et al.

19. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach. In: CVPR. pp. 2083–2090 (2013) 3
20. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: CVPR (June 2015) 1
21. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of intelligence, pp. 115–141. Springer (1987) 2
22. Lee, H., Kim, D.: Salient region-based online object tracking. In: WACV. pp. 1170–1177. IEEE (2018) 1
23. Li, G., Xie, Y., Wei, T., Wang, K., Lin, L.: Flow guided recurrent neural encoder for video salient object detection. In: ICCV. pp. 3243–3252 (2018) 2, 3, 11
24. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR. pp. 5455–5463 (2015) 3
25. Li, H., Chen, G., Li, G., Yu, Y.: Motion guided attention for video salient object detection. In: ICCV (2019) 11
26. Li, S., Seybold, B., Vorobyov, A., Lei, X., Jay Kuo, C.C.: Unsupervised video object segmentation with motion-based bilateral networks. In: ECCV. pp. 207–223 (2018) 3
27. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: CVPR (2019) 1, 3
28. Liu, N., Han, J., Yang, M.H.: Picanet: Learning pixel-wise contextual attention for saliency detection. In: CVPR. pp. 3089–3098 (2018) 3
29. Liu, P., Lyu, M., King, I., Xu, J.: Selflow: Self-supervised learning of optical flow. In: CVPR. pp. 4571–4580 (2019) 7
30. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015) 3
31. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: CVPR (2019) 4
32. Mechrez, R., Shechtman, E., Zelnik-Manor, L.: Saliency driven image manipulation. Machine Vision and Applications **30**(2), 189–202 (2019) 1
33. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV. pp. 1777–1784 (2013) 3
34. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: CVPR. pp. 733–740 (2012) 10
35. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR. pp. 724–732 (2016) 6, 10
36. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: CVPR. pp. 7479–7489 (2019) 1, 2, 8, 11
37. Rahtu, E., Kannala, J., Salo, M., Heikkilä, J.: Segmenting salient objects from images and videos. In: ECCV. pp. 366–379 (2010) 3, 11, 14
38. Shafieyan, F., Karimi, N., Mirmahboub, B., Samavi, S., Shirani, S.: Image seam carving using depth assisted saliency map. In: ICIP. pp. 1155–1159. IEEE (2014) 1
39. SHI, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.k., WOO, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: NeurIPS. pp. 802–810 (2015) 4
40. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convlstm for video salient object detection. In: ECCV. pp. 715–731 (2018) 2, 4, 11, 14
41. Squire, L.R., Dronkers, N., Baldo, J.: Encyclopedia of neuroscience. Elsevier (2009) 2
42. Tang, Y., Zou, W., Jin, Z., Chen, Y., Hua, Y., Li, X.: Weakly supervised salient object detection with spatiotemporal cascade neural networks. IEEE Transactions on Circuits and Systems for Video Technology (2018) 11

43. Tu, W.C., He, S., Yang, Q., Chien, S.Y.: Real-time salient object detection with a minimum spanning tree. In: CVPR. pp. 2334–2342 (2016) 3, 11

44. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. IJCV **103**(1), 60–79 (2013) 1

45. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV. pp. 3551–3558 (2013) 1

46. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. In: CVPR. pp. 4305–4314 (2015) 1

47. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting video saliency: A large-scale benchmark and a new model. In: CVPR (2018)

48. Wang, W., Shen, J., Shao, L.: Consistent video saliency using local gradient flow optimization and global refinement. IEEE TIP **24**(11), 4185–4196 (2015) 3, 10

49. Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. IEEE TIP **27**(1), 38–49 (2017) 3, 11, 14

50. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C.H., Ling, H.: Learning unsupervised video object segmentation through visual attention. In: CVPR (June 2019) 4

51. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004) 8

52. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR. pp. 3166–3173 (2013) 10

53. Yang, J., Yang, M.H.: Top-down visual saliency via joint crf and dictionary learning. IEEE TPAMI **39**(3), 576–588 (2016) 3

54. Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., Li, S.Z.: Salient color names for person re-identification. In: ECCV. pp. 536–551. Springer (2014) 1

55. Yang, Z., Wang, Q., Bertinetto, L., Bai, S., Hu, W., Torr, P.H.: Anchor diffusion for unsupervised video object segmentation. In: ICCV (2019) 4

56. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: Unitbox: An advanced object detection network. In: ACM MM. pp. 516–520 (2016) 8

57. Zhang, L., Dai, J., Lu, H., He, Y., Wang, G.: A bi-directional message passing model for salient object detection. In: CVPR. pp. 1741–1750 (2018) 11, 14

58. Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G.: Progressive attention guided recurrent network for salient object detection. In: CVPR (June 2018) 1

59. Zhang, Z., Qiao, S., Xie, C., Shen, W., Wang, B., Yuille, A.L.: Single-shot object detection with enriched semantics. In: CVPR. pp. 5813–5821 (2018) 4

60. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: CVPR. pp. 1265–1274 (2015) 3

61. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR. pp. 3586–3593 (2013) 1

62. Zhao, R., Oyang, W., Wang, X.: Person re-identification by saliency learning. IEEE TPAMI **39**(2), 356–370 (2016) 1

63. Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. In: CVPR (2019) 3