# Progressively real-time video salient object detection via cascaded fully convolutional networks with motion attention

Zheng, Qingping; Li, Ying; Zheng, Ling; Shen, Qiang

*Published in:*
Neurocomputing

*DOI:*
10.1016/j.neucom.2021.10.007

*Publication date:*
2022

*Citation for published version (APA):*
Zheng, Q., Li, Y., Zheng, L., & Shen, Q. (2022). Progressively real-time video salient object detection via cascaded fully convolutional networks with motion attention. *Neurocomputing*, *467*, 465-475. https://doi.org/10.1016/j.neucom.2021.10.007

# Progressively Real-time Video Salient Object Detection via Cascaded Fully Convolutional Networks with Motion Attention

Qingping Zheng[a], Ying Li[b,*], Ling Zheng[c] and Qiang Shen[d]

[a]School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, 710072, China
[b]School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, 710072, China
[c] School of Informatics, Xiamen University, Xiamen,361005, China
[d]Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, U.K.

ARTICLE INFO

*Keywords*:
Video salient object detection, cascaded fully convolutional networks, motion attention, optical flow

ABSTRACT

Semantics and motion are two cues of essence for the success in video salient object detection. Most existing deep-learning based approaches extract semantic features by the use of only one fully convolutional network with a simple stacked encoders. They simulate motion patterns of video objects with two consecutive frames being simultaneously fed into a convolutional LSTM network or a weights-sharing fully convolutional network. However, such approaches have the shortcomings of producing a coarse predicted saliency map or requiring significant computational overheads. In this paper, we present a novel approach with cascaded fully convolutional networks involving motion attention (abbreviated as CFCN-MA), to achieve real-time saliency detection in videos. Our key idea is to construct twofold fully convolutional networks in order to gain a saliency map from coarse to fine. We devise an optical flow-based motion attention mechanism to improve the prediction accuracy of the initial fully convolutional networks, using the popular FlowNet2-SD model that is efficient and effective for motion pattern recognition of distinctive objects in videos. This method can obtain a fine saliency map with a refined region of interest. Moreover, we propose a means for calculating attention-guided intersection-over-union loss (shortnamed as *AIoU*) to supervise the CFCN-MA model in learning a saliency map with both clear edge and complete structure. Our approach is evaluated on three popular benchmark datasets, namely DAVIS, ViSal and FBMS. Experimental results demonstrate that our method outperforms many state-of-the-art techniques while meeting the real-time demand at 27 fps.

## 1. Introduction

Salient object detection aims to identify regions of interest from images and videos. This can serve as a prepossessing method for many other application problems in both video analysis and image analysis, such as scene understanding [42], visual tracking [2], and person re-identification [46]. The saliency detection can be roughly divided into two types of task, namely human eye fixation prediction and salient object detection. The slight difference between them is that the former targets at distinguishing the fixation points at first glance and the latter at segmenting the obvious objects in scenes. In the area of image modelling and analysis, the task of salient object detection, highly correlated to semantic segmentation, has rekindled extensive studies since the fully convolutional network (FCN) [24] was proposed. In this paper, we focus our attention on the problem of salient object detection in videos.

Video salient object detection is more challenging than image salient object detection, since objects in videos are not only semantically relevant but also temporally relevant. Video objects may be dynamically changing and the region of interest in a video sequence may suffer from a constant variation over time, including: deformation of different degree, transformation in colour and variation on scales. Unlike image saliency detection where semantic clue has a decisive impact on the prediction of results, motion information between two consecutive frames plays a significant role in video saliency detection as human viewer is prone to paying a higher attention on objects of faster movement. In addition, motion patterns of video objects can work as an auxiliary cue to facilitate the detection of certain prominent regions whose appearance may change constantly as time goes by and may seem to be very similar to that of the cluttered background. Nevertheless, how to effectively integrate semantic cue and motion cue remains a critical challenge in the literature.

Existing approaches for detecting video salient objects mainly involve two steps: first to extract the spatial and temporal features, and then to apply a spatio-temporal fusion strategy to produce a final saliency map. In particular, deep learning models have been shown to offer a substantially higher accuracy than traditional methods, due to their strong capability of feature representation. Typically, advance in deep-learning based video salient object detectors has been driven by the use of fully convolutional networks [24] and convolutional LSTM (convLSTM) [41], for semantic feature (or high-level spatial feature) extraction and motion feature (or temporal feature) extraction, respectively. For instance, the FCN-based model as proposed by Wang *et al.* [36] exploits one FCN with its input being an image for semantic extraction and another weights-sharing FCN with two consecutive frames concatenated together to act as the input for motion extraction. This method has achieved an impressive performance in detection accuracy but suffers from the problem of coarse prediction, due to the loss of many details in high-level spatial features. The convLSTM

*Corresponding author.
lybyp@nwpu.edu.cn (Y. Li)

model proposed by Shi *et al.* [41] entails an excellent performance on edge detection. The prediction accuracy of the resulting deep-learning models for video salient object detection is considerably improved through taking advantage of FCN for semantic feature extraction, followed by employing a bi-directional convLSTM for spatiotemporal feature fusion [20, 21, 29]. However, the coarse detection problem remains: the approach has the difficulty in detecting small salient objects. This is largely due to the fact that the fusion of spatial and temporal features is carried out at a rather late stage during the overall detection process.

To address this problem, we present a method, shortnamed CFCN-MA in this paper. It works by initially constructing a semantic FCN for the prediction of a coarse saliency map, while utilising a FlowNet model to extract motion features and fusing such features with a motion attention module (to enhance coarse saliency detection), and ultimately by leveraging another cascaded FCN to obtain a refined final detection outcome. It combines two key ideas: One takes advantage of cascaded fully convolutional networks for obtaining the semantic features from coarse to fine, and the other implements the strategy of optical flow-based knowledge transfer learning for effective extraction of motion information, creating a motion based channel attention to rectify coarse semantic features.

Compared to the existing FCN-based approach (e.g., [36]), CFCN-MA offers a "coarse-to-fine" framework, with two sub-networks cascaded to resolve the coarse prediction problem that would otherwise result from the use of only one fully convolutional network involving a simple stack of convolution layers. Note that previous FCN-based models encode feature hierarchies in a non-linear local-to-global pyramid, causing deeper semantic features to be coarser due to the loss of further low-level spatial details. To address this important shortcoming, each fully convolutional network in CFCN-MA incorporates features obtained from multiple layers into the computation of the final result, directly or indirectly.

This general design works well for simple scenes, but may fail to separate a region of interest from certain complicated scenes. For example, background context may be almost the same as the appearance of salient objects, or the region of interest occupies quite a small proportion within the whole frame. Fortunately, it is possible to address these issues in videos by taking the temporal information into consideration. Previous deep models attempt to exploit a sequential structure of an FCN followed by a convolutional LSTM framework (referred to as FCN-ConvLSTM hereinafter), to fuse spatial and temporal features. This is not sufficiently efficient to achieve comprehensive spatiotemporal features due to the late incorporation of temporal information. However, in dealing with a dynamic visual scene, optical flow, regarded as a motion pattern of object surface and edges, can be utilised to detect small moving objects. Inspired by this observation, CFCN-MA employs optical flow as temporal information, thereby achieving the fusion of spatiotemporal characteristics with an attention mechanism. Considering the lack of labeled optical flow information in the problem domains concerned, a pre-trained optical flow model is herein use to extract the motion features.

In practice, real-time video salient object detection is often required, leading to the challenge of trading off between accuracy and real-time performance. For this purpose, we intend to reduce the amount of network parameters as much as possible, while maintaining the insurance regarding accuracy. Consequently, in devising the present approach, within the first semantic fully convolutional network, the structure is set to contain only a small backbone, to be followed by a lightweight refinement fully convolutional network in a cascaded manner. Between them, a motion attention module is designed that employs an optical flow model named FlowNet2-SD, in an effort to ensure a better trade-off between computational efficiency and accuracy.

Our contributions are threefold: (1) Development of a cascaded fully convolutional network system, including a semantic fully convolutional network, which is utilised to capture the spatial context of static images in order to obtain a coarse saliency map, and another lightweight refinement fully convolutional network, to further obtain a final fine saliency map. (2) Design of a motion attention module by adopting optical flow-based motion information, to generate an enhanced saliency map with an efficient pre-trained FlowNet2-SD model, which helps deal with small displacements while performing optical flow extraction, to satisfy real-time requirement. (3) Proposal for a method of computing attention-guided intersection-over-union (*AIoU*) loss, which is exploited to reduce the representation lose of any internal structure within salient objects, while focusing on edge learning.

The rest of this paper is organized as follows. Section 2 presents an overview of related work to the developments reported herein. Section 3 details the proposed approach. Section 4 shows experimental results and finally, Section 5 concludes this work and points out directions for interesting further research.

## 2. Related Work

### 2.1. Models for Video Salient Object Detection

Saliency detection can be classified into human eye fixation prediction and salient object detection, involving saliency detection and analysis in images [33] or in videos [11, 12, 13, 37]. The main difference between human eye fixation and salient object detection is: The former aims to predict the distribution of human fixation points, whereas the latter does to perform binary classification for each pixel in an image or a single video frame. Over the past two decades, saliency detection in images has been intensively studied, while video saliency detection is still a relatively unexplored territory. In this paper, we focus on the work of highlighting the main salient objects in videos, that is, the work on video salient object detection.

1) Conventional Models: Most previous investigations (e.g., SGSP [22], SPVM [23], SAGM [34], and GFVM [35]) of video salient object detection are simple extensions of

existing image salient object models, while assuming certain additional motion features. Such models exploit both hand-crafted spatial features in a bottom-up mechanism and temporal information (e.g., optical flow and difference-over-time), with limited representation ability.

2) Deep-learning Models: Deep learning-based models for video salient object detection have also been proposed. Performance-wise, these models beat the traditional methods by a large margin, benefited from large-scale datasets and the strong learning ability of deep neural networks. Typically, early deep learning models applied in video salient object detection are devised for spatial feature extraction. With a great success of utilising a fully convolutional network (FCN) for image segmentation [24], the DLVS model [36] exploits the FCN structure to predict salient object in videos, also achieving a promising performance. Subsequently, a number of fully convolutional networks combined with convLSTM (FCN-ConvLSTM), such as FGRN [20], PDBM [29] and SSAV, have been put forward for video salient object prediction.

## 2.2. Related Network Design

1) Cascaded FCN Structure: Currently, deep learning based models for video salient object detection [7, 16, 20, 29] exploit either FCN structure or FCN-ConvLSTM structure. The FCN structure-based approach only includes one fully convolutional network. For example, the DLVS model [36] utilises a fully convolutional network for the detection of static salient objects (in a single video frame), and then feeds two concatenated adjacent frames into the same fully convolutional network to predict salient objects in dynamic scenes. Such early use of the FCN structure is composed of a stack of convolution operations (also known as down-sampling), resulting in the potential severe loss of low-level spatial information and hence, often yielding a coarse inference outcome. Another type of FCN-ConvLSTM structure based model (e.g., FGRN [20], PDBM [29], and SSAV [11]) works by combining a fully convolutional network and a recurrent network to exploit both spatial feature and temporal information, in implementing video salient object prediction. However, the spatial and temporal features are only incorporated together in a sequential manner, failing to learn spatiotemporal features simultaneously and comprehensively.

This paper extends the FCN structure to a cascaded FCN structure, with the employment of two FCNs, named semantic network and refinement network respectively. In order to obtain a high-resolution feature map, each FCN takes advantage of its inherent encoder-decoder architecture with a skip-connected mechanism, to integrate the deep, low-resolution feature maps in support of more accurate object detection.

2) Motion Attention Mechanism: Attention mechanisms [11, 19, 38] have been widely used in video salient object detection. However, for motion based attention mechanism, how to effectively represent motion information between two adjacent frames remains a significant challenge. Optical flow can be regarded as a means to depict the motion of individual pixels on a given image plane, offering a principled method to compute the motion of image intensities in the scene under consideration. Early video saliency detection techniques mainly employ the conventional Lucas-Kanade mechanism [3] or its variant [4] to compute optical flow, thereby being not sufficiently accurate while requiring heavy computation. Recently, Dosovitskiy *et al.* [9] utilised a convolutional neural network (known as FlowNet) to model optical flow with high accuracy, but its speed remains unsatisfactory for real-time applications.

Through further optimisation of FlowNet, Eddy *et al.* [15] proposed FlowNet2.0, which has achieved the best performance on both accuracy and speed so far. In FlowNet2.0, several sub-system components (of a different number of parameters) are introduced to deal with various motion characteristics, including FlowNet2-S (147M weights), FlowNet2-C (149M weights) and FlowNet2-SD (173M weights). Amongst them, FlowNet2-SD is shown to be able to cope with small displacements, and FlowNet2-C is able to compute optical flow with large displacements. Moreover, FlowNet2-SD has a better performance than FlowNet2-C on dealing with small objects. This property can be utilised to resolve the aforementioned problem when facing the situations where salient objects are too small, or when objects of interest may be occluded by other non-salient objects. Following this idea, we employ FlowNet2-SD to simulate the temporal information between two continuous frames in the present work. Different from the composite attention method as introduced by Lai *et al.* [19], our method forwards the obtained optical flow map to a self-attention mechanism, enabling an optical flow based saliency map to be computed. The resulting saliency map is then multiplied by the coarse saliency map that has been previously generated from the semantic FCN via dot product, to achieve a spatiotemporal-fused saliency map.

# 3. Proposed Approach

Figure 1 illustrates the general framework of our proposed system, comprising cascaded fully convolutional networks with motion attention, for progressively real-time video salient object detection. Given two adjacent frames in a video sequence, the semantic FCN is firstly employed for coarse saliency detection of a single (current) frame. Then, the resulting two consecutive frames are simultaneously fed into the pre-trained optical flow model (FlowNet2-SD) to produce motion features. The motion attention module exploits such motion features to enhance the saliency map obtained from the semantic FCN, resulting in a spatiotemporal based coarse saliency map. Finally, the refinement FCN is used to refine it to obtain a final saliency map. In the following subsections, we describe the details of how to extract a coarse saliency map, and of how to integrate the motion priors into the coarse-to-fine procedure for fine saliency detection.

## 3.1. Coarse Saliency Map via Semantic FCN

The variant U-shaped fully convolutional network [28] is herein taken to implement the semantic FCN, to predict a

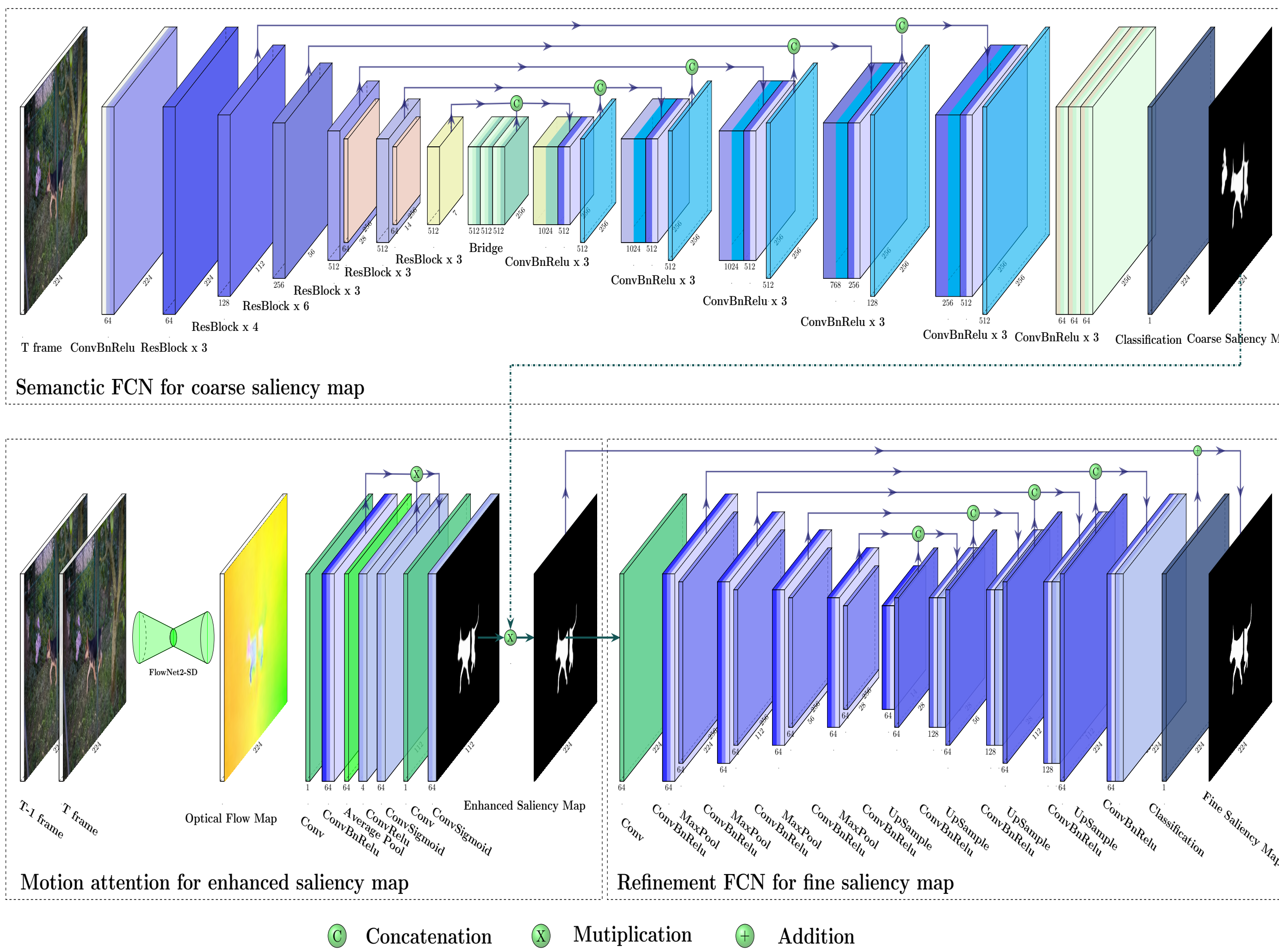



**Figure 1:** Proposed CFCN-MA framework is comprised of three parts: a semantic fully convolutional network for predicting a coarse saliency map, a motion attention module for generating an enhanced saliency map, and a refinement fully convolutional network for yielding an ultimately fine saliency map.

coarse saliency map for each frame in a given video clip. This choice is based on the observation that the original fully convolutional network and its variants have been extensively studied for semantic segmentation or image salient object detection, capable of achieving breakthrough results. In videos, salient object detection can be treated as a binary segmentation problem which simply separates the region of interest from a clustered background. The U-shaped fully convolutional network is the state-of-the-art method for border region segmentation. At the stage of coarse prediction, we employ one bit deeper-and-wider U-shaped semantic FCN with ResNet34 as the backbone. Here, by "deeper-and-wider" it means a deeper depth and wider channel [14, 43]. The reason behind this design decision is that a larger model can prevent under-fitting given sufficient training data.

As shown in Figure 1, the semantic FCN contains seven encoder layers, one bridge layer, six decoder layers and one classifier. Within the encoder part, the first encoder layer is a fundamental convolution unit (ConvBnRelu), composed of a convolution operation, a batch normalisation and a rectified linear (ReLU) activation function. The second-to-fifth encoder layers correspond to the first-to-forth layers of ResNet34, respectively. The sixth and seventh encoder layers each involve three stacked residual blocks (ResBlocks). Note that the fifth and sixth encoder layers are each followed by a max-pooling operation. The bridge part contains three stacked fundamental convolution units. Within the decoder part, except the last decoder layer, each layer has three fundamental convolution units, cascading the output of the preceding layer and that of the corresponding encoder layer together before proceeding to the first convolution unit. This skip-connected mechanism effectively integrates low-level features from multi-layers into high-level semantic features, thereby improving the accuracy of salient object detection. Finally, a convolution followed with a sigmoid activation function is used as the classifier.

Given each frame of a video sequence, a coarse saliency map is obtained by passing it onto this semantic FCN. In order to speed up the model convergence in the training phase, each decoder layer is followed by the corresponding classifier and guided by a loss. In recognition of the practical limit of having (relatively) scarce training data for video saliency detection, a dataset acquired from static image object detection is employed for the training of this semantic FCN to obtain a coarse saliency map.

## 3.2. Motion Attention for Enhanced Saliency Map

The essential difference between saliency detection in images and that in videos is that video objects are dynamic and the appearance of the same salient objects may change constantly, whereas image saliency detection does not have this problem. A semantic FCN may perform well in either image or video domains, especially when the appearance of the detected objects forms in great contrast against the background. Nonetheless, it may fail in certain cases, where salient objects have similar appearance features to those of the background, e.g., of the same illumination, indistinguishable colour and overlapped texture, or where they become too small in size over the time Specific motion cues in a video sequence are therefore employed to tackle this issue.

Optical flow offers an effective representation for relative motion patterns of the object surface and edges from one frame to the next. In our method, we apply optical flow features in the design of motion attention in an effort to improve the prediction accuracy. However, the (current) data available for video salient object detection does not have any annotated information of optical flow, so it cannot be utilised to train an optical flow model from scratch. Inspired by the work on specific-domain knowledge based transfer learning [5], the existing optical flow model is adapted to extract motion features between two adjacent frames. Considering the real-time demand for video related applications, the selected optical flow model must have a high accuracy whilst not being too time-consuming. Here, FlowNet2-SD [15], which has a good trade-off between accuracy and running speed, is chosen to extract the motion feature set $\mathbf{M}$. To fuse the motion priors within the proposed framework, motion prior is introduced based attention to enhance the coarse saliency map $\mathbf{C}$ obtained from the semantic FCN.

This motion attention strategy performs a convolution operation followed by another convolution unit in order to deal with the extracted motion map and therefore, the resulting map $\mathbf{F}$. This is further handled by an adaptive average pooling and two convolutions which are subsequently, followed by ReLU and a sigmoid function. We denote the output of this channel attention as $\mathbf{A}$. Consequently, the enhanced motion map $\mathbf{M}'$ can be generated and formulated such that

$$\mathbf{M}' = \mathbf{A} \times \mathbf{F} \tag{1}$$

where the range of $A$ is from 0 to 1, and the value of $M'$ belongs to $(0, +\infty)$. Finally, the enhanced motion map $\mathbf{M}'$ is passed onto the classification module (a convolution followed by a sigmoid function) to obtain a temporal saliency map $\mathbf{T}$, which is subsequently employed as the motion attention cue to correct the coarse saliency map $\mathbf{C}$ returned by the semantic FCN, through

$$\mathbf{C}' = \mathbf{C} \times \mathbf{T} \tag{2}$$

where $\mathbf{C}'$ is the resulting enhanced saliency map, and the value range of the motion attention map $\mathbf{T}$ is [0, 1].

## 3.3. Fine Saliency Map via Refinement FCN

Although optical flow based motion patterns are quite useful for video salient object detection, they may correspond to immobile objects which are not salient, due to camera motion [21]. Such introduced noise (non-salient objects) makes the enhanced saliency map $\mathbf{C}'$ become worse than the previous obtained coarse saliency map. To address this problem, another V-shaped fully convolutional network, termed refinement FCN hereafter, is further introduced to make the enhanced saliency map $\mathbf{C}'$ more accurate. Since the amount of training data for video saliency detection is typically inadequate, at this fine-tuning stage, a lightweight residual refinement FCN is developed to avoid over-fitting.

The structure of this refinement FCN is comprised of four encoders, four decoders and one classifier. Every encoder except the first one contains a convolution unit (ConvBnRelu) and a max-pooling operation. The first encoder adds an extra convolution operation before the convolution unit in order to deal with the enhanced motion map $\mathbf{M}'$. Each decoder consists of a basic convolution unit and an up-sampling operation. Before decoding, the previous up-sampling result and the corresponding convolution result are cascaded together. The classifier implements a simple convolution operation. The final saliency map $\mathbf{S}_{final}$ is obtained by directly adding the fine saliency map $\mathbf{S}_{fine}$ onto the enhanced saliency map $\mathbf{C}'$, as formulated by

$$\mathbf{S}_{final} = \mathbf{C}' + \mathbf{S}_{fine} \tag{3}$$

## 3.4. Training Loss

As mentioned previously, saliency detection is of high relevance to segmentation. More specifically, saliency detection can be treated as a binary (foreground and background) classification problem at pixel level. Therefore, the binary cross entropy (*bce*) loss [8] can be used to train the entire model. However, the *bce* loss does not consider any relationship among pixels, and a fully convolution network may suffer from the problem of coarse prediction, especially concerning blur edge and structure loss.

To tackle this problem, while noticing the success of utilising the generalised intersection-over-union (*GIoU*) loss [27] for bounding box regression, we propose an attention-guided intersection-over-union (*AIoU*) loss, including an *IoU* component to clear the edge of salient objects and an attention-guided component to retrieve internal structure of salient objects. Figure 2 illustrates the idea of this proposed *AIoU* loss. In particular, Figure 2(a) shows an example of a predicted saliency map guided by the *bce* loss. The yellow area in Figure 2(b) illustrates the predicted saliency map supervised by the combination of the *bce* loss and the *IoU* loss. The contour labeled in red line in this figure depicts the edge of the saliency map. It can be seen that the result predicted via the combination of the *bce* loss and the *IoU*

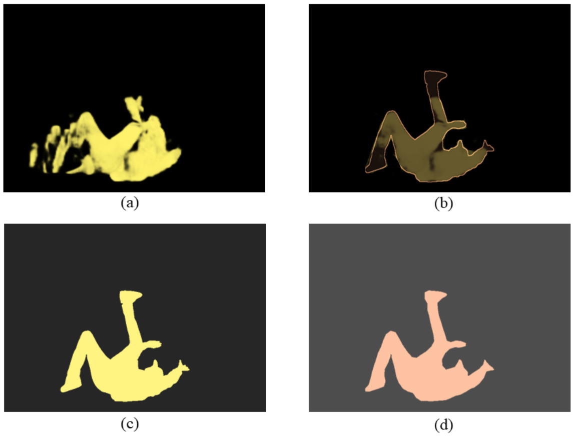



**Figure 2:** Results of saliency map predicted using different losses: (a) is a saliency map supervised by *bce* loss; (b) and (c) are saliency map guided by *bce* loss combined with *IoU* loss and that by proposed *AIoU* loss, respectively; (d) is ground truth of corresponding saliency map.

loss is of good quality on the edges. However, this combination is unable to sufficiently preserve the structure of the underlying salient object, because it only considers true positive/negative pixels but ignores the false negative/positive ones. To improve the sensitivity to the structure of the salient object, the *IoU* loss is herein modified as the attention-guided (*AIoU*) loss, paying an extra attention on any incorrectly predicted pixels within the salient region.

The proposed *AIoU* loss can be formulated as

$$l_{AIoU} = 1 - (IoU - \frac{\mathbf{E}}{\mathbf{G}}) \tag{4}$$

where $\mathbf{E}$ denotes the number of misclassified pixels that should belong to the foreground but predicted as the background, and $\mathbf{G}$ is the sum of a binary ground truth pixels. As the *IoU* loss is an intersection-over-union, it makes an effective guidance in the learning of the clear edge of salient objects, but may cause certain inner structure loss. To enhance the learning of the salient object structure, we add a correction term $\frac{\mathbf{E}}{\mathbf{G}}$ (to the *IoU*), representing an error rate of the pixels mis-predicted within a salient region. In so doing if there are more salient pixels being wrongly predicted, the value of the total loss will increase.

The overall cost function is therefore,

$$L = \sum_{(k=1)}^{K} l_{bce}^{(k)} + l_{AIoU}^{(k)} \tag{5}$$

where $l^{(k)}$ is the $k^{th}$ sample loss, $K$ is the number of frames in the video clip addressed, $l_{bce}^{(k)}$ denotes the *bce* loss, and $l_{AIoU}^{(k)}$ denotes the proposed attention-guided *IoU* loss. In summary, to speed up the convergence of the entire model, we add this loss to the end of each decoder part at the refinement stage.

# 4. Results and Discussions

## 4.1. Datasets

Our proposed approach is evaluated on four popular public benchmark datasets: Densely Annotated Video Segmentation (DAVIS) [45], Freiburg-Berkeley Motion Segmentation (FBMS) [42], ViSal [17], and DAVSOD [11]. The DAVIS dataset is originally built for video object segmentation. It has 50 high-quality video sequences, covering different technical challenges, such as occlusions, motion-blur and appearance changes. The FBMS dataset is initially created for motion segmentation, covering 59 video sequences with a split into a training set (29 video sequences) and a test set (30 video sequences). In this dataset, there are multiple objects moving at the same time. The ViSal dataset is the earliest for video salient object detection, collected from the existing video datasets and YouTube, including 17 video sequences with 963 frames and 193 annotated frames in total. DAVSOD is the largest scale dataset for video salient object detection, including 90 training, 46 validation and 90 testing (split into 35 easy, 30 normal and 25 difficult) videos. The performances of our model and other alternatives are compared on the DAVIS test set, the FBMS test set, the entire ViSal dataset (because there is no split of testing and training sets in the ViSal dataset) and the DAVSOD-35 dataset.

## 4.2. Evaluation Metrics

There are three widely-used performance measures in video saliency detection, including: mean absolute error (MAE) $\mathcal{M}$ [25], F-measure $\mathcal{F}$ [1], and S-measure $\mathcal{S}$ [10].

Given a saliency map $\mathbf{S}$, it initially has to be converted into a binary mask. Then precision and recall can be defined respectively as below:

$$Precision = \frac{|\mathbf{S} \cap \mathbf{G}|}{|\mathbf{S}|} \tag{6}$$

$$Recall = \frac{|\mathbf{S} \cap \mathbf{G}|}{|\mathbf{G}|} \tag{7}$$

where $|\cdot|$ stands for the number of non-zero binary pixels, and $\mathbf{G}$ denotes the collection of binary ground-truth pixels.

The MAE metric considers both salient and non-salient pixels. It calculates the average difference between a final saliency map $\mathbf{S}$ and a binary ground-truth $\mathbf{G}$, such that

$$MAE = \frac{1}{(wh)} \sum_{x=1}^{w} \sum_{y=1}^{h} ||\mathbf{S}(x,y) - \mathbf{G}(x,y)|| \tag{8}$$

where $w$ and $h$ are the width and the height of an input frame respectively, and both saliency map $\mathbf{S}$ and ground truth $\mathbf{G}$ are normalised to the values between 0 and 1.

The F-measure is a weighted harmonic of precision and recall, defined as

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall} \tag{9}$$

where $\beta^2 = 0.3$ is assigned to allocate more weight to precision than recall. A set of F-measure values is first computed for each saliency map with the threshold ranging from

0 to 255, leading to an average F-measure score. Then, a sequence of such mean F-measure scores is computed with respect to all predicted saliency maps, with the maximum mean F-measure selected as a final evaluation index.

The S-measure captures the similarity over non-binary foreground maps, comprising a region-aware structural similarity and an object-aware structural similarity, which is defined by

$$S = \alpha \cdot S_o + (1 - \alpha) \cdot S_r \quad (10)$$

where $\alpha \in [0, 1]$ and $\alpha$ is herein empirically set to 0.5. The further details of the computation of $S_o$ and $S_r$ are omitted here but can be found in [10].

## 4.3. Experimental Setup and Implementation

At the training phase, due to the inadequate amount of the training data for video saliency detection, we adopt the largest image salient dataset (DUTS-TR) [32], to train the semantic FCN first. This dataset contains many diverse salient objects, and totally has 10533 images. To train CFCN-MA, the above pre-trained semantic FCN in the static image domain is used to initialise the weights of the semantic FCN component within it and the pre-trained FlowNet2-SD model is used to obtain the optical flow between two adjacent frames. Then, for testing on the DAVIS, FBMS and ViSal datasets, we combine the DAVIS and FBMS training datasets to train the entire CFCN-MA model end-to-end. For testing over the DAVSOD dataset, we use the DAVSOD training dataset to train CFCN-MA. Also, each image is first re-scaled to 256 x 256 and then resized to 224 x 224 via a bilinear interpolation. The entire model is optimised using the Adam optimiser [18], with a learning rate of 0.001 and other default hyper parameters typically used in the literature. We train the CFCN-MA network for approximately 100K iterations.

Our proposed method is implemented on the commonly-used open source framework: Pytorch 0.4.1. A 16-core PC with an Intel(R) Xeon(R) E5-2620 v4 2.10GHz CPU (with 512 GB RAM) and four GeForce GTX 1080 Ti GPUs (with 11GB memory) are used to train and test the model. The total size of our proposed CFCN-MA is 260M, including the 173M FlowNet2-SD and 87M remaining modules.

## 4.4. Comparison with State-of-the-arts

We quantitatively and qualitatively compare the proposed approach with other 17 methods, including ten traditional approaches (SIVM [26], TIMP [47], SPVM [23], RWRV [17], MB+M [44], SAGM [34], GFVM [35], STBP [40], SGSP [22], SFLR [39]), and seven deep-learning based approaches (MSTM [31], SCOM [6], SCNN [30], DLVS [36], FGRN [20], MBNM [21], and PDBM [29]).

### 4.4.1. Quantitative Evaluation

Table 1 shows the results of quantitative comparison between our method (CFCN-MA) and other competing approaches, on four datasets in terms of all evaluation metrics (namely (MAE) $\mathcal{M}$, F-measure $\mathcal{F}$, and S-measure $\mathcal{S}$. It demonstrates that deep learning-based methods for video saliency detection significantly surpass the classical methods. CFCN-MA is also a deep learning-based approach, and its performance is superior to all others, across all four datasets regarding almost all evaluation metrics. In particular, for MAE, F-measure and S-measure, our method almost ranks the top on all test datasets.

Examining these results more closely, we have the following noteworthy observations: (1) Deep learning based methods consistently outperform conventional methods by a large margin. Different from the conventional saliency detection methods which mainly rely on man-made features, deep learning based methods can generate features automatically. This further verifies that deep features beat human-made features on video salient object detection. (2) Our method is of the lowest MAE value and the highest F-measure and S-measure values amongst all deep-learning based methods on all datasets. Particularly, these results show that CFCN-MA outperforms the other FCN-based models (i.e., DLVS and PDBM). This is attributed to the proposed motion attention using the optical flow as prior knowledge, different from the approach taken by the others that simulates motion features by directly concatenating two successive frames and forwarding the combined outcomes into a simple FCN model or convLSTM model. (3) We can draw a conclusion from all these results that our method has a better generalisation ability than other methods.

### 4.4.2. Qualitative Evaluation

In order to qualitatively compare our method with the rest, Figure 3 shows representative visual examples in different challenging cases, such as small-size salient objects, region of interest occluded by other objects (of no interest), and object texture similar to background. As shown in this figure, most results lose the structure information and mis-predict many non-salient pixels as salient ones, whereas CFCN-MA offers better results. In another word, our method achieves a better visual performance than the rest, beating the previous deep models (i.e. DLVS and PDBM). This further verifies the effectiveness of utilising the proposed attention-guided *IoU* loss and motion priors of FlowNet2-SD on forecasting the movement of small objects. For instance, the salient objects in the last three columns are of a very small size. Most compared methods fail to exactly identify them, while ours successfully captures each with clearer edges and a better preserved structure. The image in the second column is easy to detect and not surprisingly, our method outperforms the others again on structure and edge. The image in the first column is challenging with significant occlusion. Many other compared methods basically fail to capture the salient object, while ours can still predict it with well-preserved structure.

## 4.5. Runtime Analysis

To speed up the experiments, we firstly use the pre-trained FlowNet2-SD to extract the optical flows between two frames offline, since the weights of FlowNet2-SD are fixed during the training of CFCN-MA. Compared to

**Table 1**
Quantitative comparison between proposed CFCN-MA and 17 existing methods on DAVIS, FBMS, ViSal and DAVSOD$_{35}$ (Easy test set) datasets. Best three results are shown in red, green and blue.

| | Method | DAVIS | | | FBMS | | | ViSal | | | DAVSOD$_{35}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $\mathcal{M}\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $\mathcal{M}\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $\mathcal{M}\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $\mathcal{M}\downarrow$ |
| Traditional | SIVM [10] | 0.450 | 0.557 | 0.212 | 0.426 | 0.545 | 0.236 | 0.522 | 0.606 | 0.197 | 0.298 | 0.486 | 0.288 |
| | TIMP [18] | 0.488 | 0.593 | 0.172 | 0.456 | 0.576 | 0.192 | 0.479 | 0.612 | 0.170 | 0.395 | 0.563 | 0.195 |
| | SPVM [23] | 0.390 | 0.592 | 0.146 | 0.330 | 0.515 | 0.209 | 0.700 | 0.724 | 0.133 | 0.358 | 0.538 | 0.202 |
| | RWRV [27] | 0.345 | 0.556 | 0.199 | 0.336 | 0.521 | 0.242 | 0.440 | 0.595 | 0.188 | 0.283 | 0.504 | 0.245 |
| | MB+M [26] | 0.470 | 0.597 | 0.177 | 0.487 | 0.609 | 0.206 | 0.692 | 0.726 | 0.129 | 0.342 | 0.538 | 0.228 |
| | SAGM [34] | 0.515 | 0.676 | 0.103 | 0.564 | 0.659 | 0.161 | 0.688 | 0.749 | 0.105 | 0.370 | 0.565 | 0.184 |
| | GFVM [35] | 0.569 | 0.687 | 0.103 | 0.571 | 0.651 | 0.160 | 0.683 | 0.749 | 0.105 | 0.334 | 0.553 | 0.167 |
| | STBP [47] | 0.544 | 0.667 | 0.096 | 0.595 | 0.627 | 0.152 | 0.622 | 0.629 | 0.163 | 0.410 | 0.568 | 0.160 |
| | SGSP [22] | 0.655 | 0.692 | 0.138 | 0.630 | 0.661 | 0.172 | 0.677 | 0.706 | 0.165 | 0.426 | 0.577 | 0.207 |
| | SFLR [44] | 0.727 | 0.790 | 0.056 | 0.660 | 0.699 | 0.117 | 0.779 | 0.814 | 0.062 | 0.478 | 0.624 | 0.132 |
| Deep Learning | MSTM [40] | 0.429 | 0.583 | 0.165 | 0.500 | 0.613 | 0.177 | 0.673 | 0.749 | 0.095 | 0.344 | 0.532 | 0.211 |
| | SCOM [39] | 0.783 | 0.832 | 0.048 | 0.500 | 0.613 | 0.177 | 0.673 | 0.749 | 0.095 | 0.464 | 0.599 | 0.220 |
| | SCNN [31] | 0.714 | 0.793 | 0.064 | 0.762 | 0.794 | 0.095 | 0.831 | 0.847 | 0.071 | 0.532 | 0.674 | 0.128 |
| | DLVS [36] | 0.708 | 0.794 | 0.061 | 0.759 | 0.794 | 0.091 | 0.852 | 0.881 | 0.048 | 0.521 | 0.657 | 0.129 |
| | FGRN [20] | 0.783 | 0.838 | 0.043 | 0.767 | 0.809 | 0.088 | 0.848 | 0.861 | 0.045 | 0.573 | 0.693 | 0.098 |
| | MBNM [21] | 0.861 | 0.887 | 0.031 | 0.816 | 0.857 | 0.047 | 0.883 | 0.898 | 0.020 | 0.520 | 0.637 | 0.159 |
| | PDBM [29] | 0.855 | 0.882 | 0.028 | 0.821 | 0.851 | 0.064 | 0.888 | 0.907 | 0.032 | 0.573 | 0.698 | 0.116 |
| | SSAV [11] | 0.861 | 0.893 | 0.028 | 0.865 | 0.879 | 0.040 | 0.939 | 0.943 | 0.020 | 0.603 | 0.724 | 0.092 |
| | CFCN-MA | 0.867 | 0.888 | 0.020 | 0.865 | 0.880 | 0.037 | 0.943 | 0.945 | 0.011 | 0.568 | 0.712 | 0.085 |

**Table 2**
Speed comparison against some representative methods. Symbols '*' and '+' denote CPU time and extra computation time of optical flow. Best three results are shown in red, green and blue, respectively.

| Method | SGSP [22] | SAGM [34] | GFVM [35] | SPVM [23] | DLVS [36] | MBNM [21] | PDBM [29] | CFCN-MA |
|---|---|---|---|---|---|---|---|---|
| Time(s) | 1.700*(+) | 0.880*(+) | 1.040*(+) | 6.050*(+) | 0.470 | 0.033 | 0.050 | 0.037 |

the complete FlowNet v2.0 which takes 0.05s to compute each optical flow frame, FlowNet2-SD (which is a part of FlowNet v2.0, aiming to obtain small displacements of image sequences) only costs around 0.021s. Afterwards, the image concatenated with the corresponding optical flow is forwarded to train the remaining part of CFCN-MA, taking almost 2 days to train 100 epochs on the DAVSOD training set.

Table 2 compares the inference time performance of our method against that of the seven deep-learning based video saliency models (namely SGSP [22], SAGM [34], GFVM [35], SPVM [23], DLVS [36], MBNM [21] and PDBM [29]). Note that as SGSP [22], SAGM [34], GFVM [35] and SPVM [23] are traditional approaches without the need for the speed-up of GPU and run on CPU, excluding the computation of optical flow using FlowNet v2.0, they are left out of this comparison. DLVS [36], MBNM [21], PDBM [29] and our method are timed on the same GPU but a different CPU (Intel(R) Xeon(R) E5-2620 v4 @2.10GHz for our method and Intel Core i7-6700 @3.4GHz for others), due to the different deep-learning frameworks adopted and the difficulty in reproducing the same experimental results given in the original references for the existing methods. Given a $224 \times 224$ frame, our model can achieve around 27 fps

**Table 3**
Results of single FCN and cascaded FCNs on four datasets. Best results are shown in bold.

| Method | DAVIS | | | FBMS | | | ViSal | | | DAVSOD$_{35}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $\mathcal{M}\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $\mathcal{M}\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $\mathcal{M}\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $\mathcal{M}\downarrow$ |
| SFCN | 0.760 | 0.826 | 0.041 | 0.779 | 0.816 | 0.070 | 0.908 | 0.920 | 0.022 | 0.452 | 0.630 | 0.143 |
| CFCN | **0.797** | **0.845** | **0.031** | **0.821** | **0.862** | **0.050** | **0.943** | **0.943** | **0.012** | **0.548** | **0.692** | **0.100** |


**Figure 3:** Visual comparison of dynamic saliency maps. Top-down: (a) Original images, (b) ground truth of salient objects, (c)-(j) detected outcomes by the proposed CFCN-MA, PDBM, MBNM, FGRN, DLVS, SCNN, SCOM, and MSTM, respectively.

(which is equivalent to 0.037 seconds per frame, including 0.021s for FlowNet2-SD and 0.016s for the remaining modules) without any pre-/post-processing. Thus, CFCN-MA is more efficient than DLVS and PDBM (and is very close to the best runtime performer, MBNM). Considering the relevant design specifications of DLVS and PDBM, the winning performance of our method can be attributed to the use of the sub-module (FlowNet2-SD) of FlowNet v2.0, achieving good performance on motion estimation whilst using the parameters as few as possible to reach the real-time requirement.

## 4.6. Ablation Experiments

For this experimental study, we compare the major components within the proposed model and provide empirical results based on different model settings and different motion priors. All of models are trained with the same data augmentation and identical hyper-parameters, as described in Section 4.3.

### *4.6.1. Effectiveness of Cascaded FCNs*

In order to verify the effectiveness of the proposed “coarse-to-fine" framework, we conduct experiments on the use of a single FCN (SFCN) and on that of cascaded FCNs (CFCN), with the results shown in Table 3 and (columns (g) and (h) of) Figure 4. These experimental results clearly demonstrate that the use of CFCN outperforms that of SFCN, reflecting the effectiveness of the cascaded FCN structure introduced in this work.

### *4.6.2. Effectiveness of Motion Attention*

To verify the effectiveness of the proposed motion attention module, we also conduct experiments on the use of CFCN (without motion attention) and that of cascaded FCNs with a motion attention module employing a pre-trained FlowNet2-C as the motion prior (shortnamed CFCN-MA$_C$).

**Table 4**
Results of models with or without motion attention. Best results are shown in bold.

| Method | DAVIS | | | FBMS | | | ViSal | | | $\text{DAVSOD}_{35}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $M\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $M\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $M\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $M\downarrow$ |
| CFCN | 0.797 | 0.845 | 0.031 | 0.821 | 0.862 | 0.050 | 0.943 | 0.943 | 0.012 | 0.548 | 0.692 | 0.100 |
| $\text{CFCN-MA}_{C}$ | 0.863 | 0.885 | 0.020 | 0.845 | 0.876 | 0.041 | 0.936 | 0.940 | 0.013 | 0.551 | 0.698 | 0.087 |
| $\text{CFCN-MA}_{S}$ | 0.855 | 0.881 | 0.020 | **0.870** | 0.878 | **0.034** | 0.938 | 0.940 | 0.012 | 0.553 | 0.701 | **0.082** |
| $\text{CFCN-MA}_{SD}$ | **0.867** | **0.888** | **0.020** | 0.865 | **0.880** | 0.037 | **0.943** | **0.945** | **0.011** | **0.568** | **0.712** | 0.085 |

**Table 5**
Results of CFCN-MA using different training losses. Best results are shown in bold.

| Method | DAVIS | | | FBMS | | | ViSal | | | $\text{DAVSOD}_{35}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $M\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $M\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $M\downarrow$ | $\mathcal{F}\uparrow$ | $\mathcal{S}\uparrow$ | $M\downarrow$ |
| $\text{CFCN-MA}_{bce}$ | 0.851 | 0.881 | 0.024 | 0.849 | 0.870 | 0.040 | 0.924 | 0.930 | 0.020 | 0.558 | 0.707 | 0.085 |
| $\text{CFCN-MA}_{bce+IoU}$ | 0.855 | 0.885 | 0.022 | 0.842 | 0.869 | 0.037 | 0.899 | 0.919 | 0.021 | 0.549 | 0.700 | 0.088 |
| $\text{CFCN-MA}_{bce+AIoU}$ | **0.867** | **0.888** | **0.020** | **0.865** | **0.880** | **0.037** | **0.943** | **0.945** | **0.011** | **0.568** | **0.712** | **0.085** |

The results are shown in Table 4, reflecting the positive effect of utilising the proposed motion attention module, since the motion information plays a notable role in achieving superior results. We further compare the effectiveness of using different pre-trained optical flow models, with the results given in Table 4, where $\text{CFCN-MA}_{S}$ and $\text{CFCN-MA}_{SD}$ (also known as CFCN-MA) denote the cascaded FCNs with motion attention based on FlowNet2-S and on FlowNet2-SD, respectively. It can be seen that $\text{CFCN-MA}_{SD}$ achieves better results than the other two.

### *4.6.3. Effectiveness of AIoU Loss*

In order to prove that the proposed attention-guided *IoU* loss is quite effective to supervise the entire network to learn a better salient region, we conduct the comparing experiments, using different losses to train the CFCN-MA. Here, the CFCN-MA trained with the *bce* loss, the combination of the *bce* and *IoU* losses, and the combination of the *bce* and the *AIoU* losses are labeled as $\text{CFCN-MA}_{bce}$, $\text{CFCN-MA}_{bce+IoU}$ and $\text{CFCN-MA}_{bce+AIoU}$ respectively. The comparing results shown in Table 5 reveal that our proposed *AIoU* loss achieves the best performance on all three public datasets, verifying the effectiveness of this loss function. Note that the CFCN-MA trained with the combination of the *bce* and *IoU* losses cannot guarantee an improved performance over the model trained by just *bce* loss. As explained in Section 3.4, the *IoU* loss can lead to a clear boundary of the predicted salient objects, but it cannot guarantee a complete internal semantic topology of the salient objects since it does not take false negative/positive pixels into consideration.

Last but not the least, Figure 4 includes additional visual examples to further reflect the benefits of the proposed approach, while also revealing its limitation. These qualitative results illustrate that only when both spatial saliency map and flow estimation fail, does the final saliency map fail.

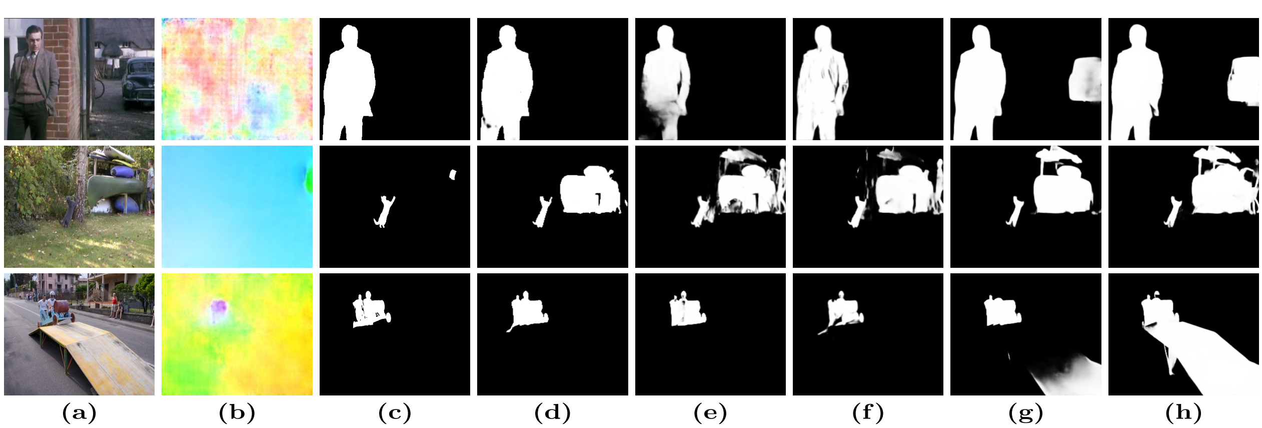
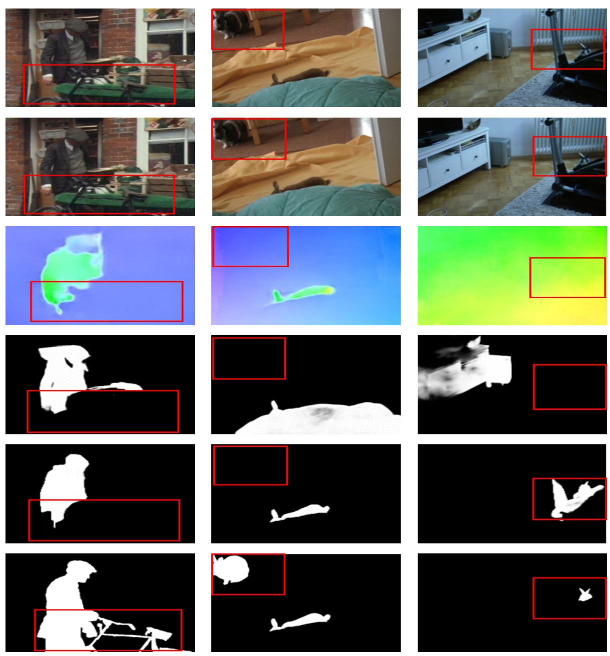



**Figure 4:** Visual comparison of dynamic saliency maps. Left-right: (a) Original images, (b) optical flow between two adjacent frames, (c) ground truth of salient objects, (d)-(h) detected outcomes by $\text{CFCN-MA}_{bce+AIoU}$, $\text{CFCN-MA}_{bce+IoU}$, $\text{CFCN-MA}_{bce}$, CFCN, and SFCN, respectively.

Otherwise, even when the motion estimation is not so accurate, provided that the coarse saliency map generated from the semantic FCN is of good quality, a highly satisfactory final saliency map can still be detected. Alternatively, if the motion estimation is accurate but the semantic FCN fails, our model can still achieve a good prediction. This verifies that our method can fuse the spatial and temporal features effectively.

### 4.7. Analysis of Failure Cases



(a) (b) (c)

**Figure 5:** Examples of failure cases. Rows from first to last correspond to previous frame, current frame, optical flow of two frames, coarse saliency map obtained from semantic FCN, final predicted saliency map, and ground truth, respectively.

Whilst the proposed method can handle most of video sequences, there are occasional cases where it fails on these datasets. Figure 5 shows the examples of such cases resulting from the application of CFCN-MA. In particular, it fails to identify the salient objects from videos when both optical flow and semantic information of the objects concerned cannot be detected correctly. There are two reasons for this. Firstly, when the context of salient objects are very similar to that of background regions, or if the sizes of salient objects are too small, the semantic FCN may fail to extract spatial features. Secondly, the inaccurate optical flow may adversely affect the robustness of temporal features. To minimise the occurrence of such failures, in further work, it is important to extract and fuse more robust spatiotemporal features.

## 5. Conclusion

In this paper, we have proposed a cascaded fully convolution network model with motion attention. It includes a semantic fully convolutional network to capture the spatial context of static images in order to obtain a coarse saliency map, and another lightweight refinement fully convolutional network to further obtain a final fine saliency map. The motion attention module exploits optical flow-based motion information to generate an enhanced saliency map, in an effort to satisfy real-time requirement. We have also presented a method that helps reduce the representation lose of any internal structure within salient objects, while focusing on edge learning. The proposed approach has been systematically evaluated against state-of-the-art alternatives, as well as against classical non-deep learning based methods, over popular datasets, demonstrating the superior performance enjoyed by our approach. For future work, it would be interesting to investigate how to ensure the extraction of only the most informative spatial and temporal features in order to improve the model efficiency, while attaining its accuracy.

## 6. Acknowledgments

We are grateful to the reviewers for their comments, which have enabled us to improve our work significantly. This work was supported in part by the National Natural Science Foundation of China under Grant 61871460, in part by the Shaanxi Provincial Key Research and Development Program of China under Grant 2020KW-003, and in part by the Strategic Partner Acceleration Award, U.K., under Grant 80761-AU201.

## References

[1] Achanta, R., Hemami, S., Estrada, F., Susstrunk, S., 2009. Frequency-tuned salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 1597–1604.

[2] Avytekin, C., Cricri, F., Aksu, E., 2018. Saliency enhanced robust visual tracking, in: Proceedings of the European Workshop on Visual Information Processing (EUVIP), IEEE. pp. 1–5.

[3] Barron, J.L., Fleet, D.J., Beauchemin, S.S., 1994. Performance of optical flow techniques. International Journal of Computer Vision (IJCV) 12, 43–77.

[4] Brox, T., Malik, J., 2010. Large displacement optical flow: Descriptor matching in variational motion estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33, 500–513.

[5] Chen, J., Lécué, F., Pan, J.Z., Horrocks, I., Chen, H., 2018a. Knowledge-based transfer learning explanation, in: Sixteenth International Conference on Principles of Knowledge Representation and Reasoning.

[6] Chen, Y., Zou, W., Tang, Y., Li, X., Xu, C., Komodakis, N., 2018b. Scom: Spatiotemporal constrained optimization for salient object detection. IEEE Transactions on Image Processing (TIP) 27, 3345–3357.

[7] Cornia, M., Baraldi, L., Serra, G., Cucchiara, R., 2018. Predicting human eye fixations via an lstm-based saliency attentive model. IEEE Transactions on Image Processing (TIP) 27, 5142–5154.

[8] De Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y., 2005. A tutorial on the cross-entropy method. Annals of Operations Research 134, 19–67.

[9] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2758–2766.

[10] Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A., 2017. Structure-measure: A new way to evaluate foreground maps, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4548–4557.

[11] Fan, D.P., Wang, W., Cheng, M.M., Shen, J., 2019. Shifting more attention to video salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8554–8564.

[12] Guo, F., Wang, W., Shen, J., Shao, L., Yang, J., Tao, D., Tang, Y.Y., 2017. Video saliency detection using object proposals. IEEE Transactions on Cybernetics 48, 3159–3170.

[13] Guo, F., Wang, W., Shen, Z., Shen, J., Shao, L., Tao, D., 2019. Motion-aware rapid video saliency detection. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) .

[14] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

[15] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2462–2470.

[16] Jiang, L., Xu, M., Liu, T., Qiao, M., Wang, Z., 2018. Deepvs: A deep learning based video saliency prediction approach, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 602–617.

[17] Kim, H., Kim, Y., Sim, J.Y., Kim, C.S., 2015. Spatiotemporal saliency detection for video sequences based on random walk with restart. IEEE Transactions on Image Processing (TIP) 24, 2552–2564.

[18] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

[19] Lai, Q., Wang, W., Sun, H., Shen, J., 2019. Video saliency prediction using spatiotemporal residual attentive networks. IEEE Transactions on Image Processing (TIP) 29, 1113–1126.

[20] Li, G., Xie, Y., Wei, T., Wang, K., Lin, L., 2018a. Flow guided recurrent neural encoder for video salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3243–3252.

[21] Li, S., Seybold, B., Vorobyov, A., Lei, X., Jay Kuo, C.C., 2018b. Unsupervised video object segmentation with motion-based bilateral networks, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 207–223.

[22] Liu, Z., Li, J., Ye, L., Sun, G., Shen, L., 2016. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) 27, 2527–2542.

[23] Liu, Z., Zhang, X., Luo, S., Le Meur, O., 2014. Superpixel-based spatiotemporal saliency detection. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) 24, 1522–1540.

[24] Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440.

[25] Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A., 2012. Saliency filters: Contrast based filtering for salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 733–740.

[26] Rahtu, E., Kannala, J., Salo, M., Heikkilä, J., 2010. Segmenting salient objects from images and videos, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer. pp. 366–379.

[27] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 658–666.

[28] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer. pp. 234–241.

[29] Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M., 2018. Pyramid dilated deeper convlstm for video salient object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 715–731.

[30] Tang, Y., Zou, W., Jin, Z., Chen, Y., Hua, Y., Li, X., 2018. Weakly supervised salient object detection with spatiotemporal cascade neural networks. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) 29, 1973–1984.

[31] Tu, W.C., He, S., Yang, Q., Chien, S.Y., 2016. Real-time salient object detection with a minimum spanning tree, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2334–2342.

[32] Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X., 2017a. Learning to detect salient objects with image-level supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 136–145.

[33] Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R., 2019a. Salient object detection in the deep learning era: An in-depth survey. arXiv preprint arXiv:1904.09146 .

[34] Wang, W., Shen, J., Porikli, F., 2015a. Saliency-aware geodesic video object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3395–3402.

[35] Wang, W., Shen, J., Shao, L., 2015b. Consistent video saliency using local gradient flow optimization and global refinement. IEEE Transactions on Image Processing (TIP) 24, 4185–4196.

[36] Wang, W., Shen, J., Shao, L., 2017b. Video salient object detection via fully convolutional networks. IEEE Transactions on Image Processing (TIP) 27, 38–49.

[37] Wang, W., Shen, J., Xie, J., Cheng, M.M., Ling, H., Borji, A., 2019b. Revisiting video saliency prediction in the deep learning era. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) .

[38] Wang, W., Zhao, S., Shen, J., Hoi, S.C.H., Borji, A., 2019c. Salient object detection with pyramid attention and salient edges, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[39] Wu, Z., Su, L., Huang, Q., Wu, B., Li, J., Li, G., 2016. Video saliency prediction with optimized optical flow and gravity center bias, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), IEEE. pp. 1–6.

[40] Xi, T., Zhao, W., Wang, H., Lin, W., 2016. Salient object detection with spatiotemporal background priors for video. IEEE Transactions on Image Processing (TIP) 26, 3425–3436.

[41] Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c., 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: Proceedings of the advances in Neural Information Processing Systems (NIPS), pp. 802–810.

[42] Yuan, Y., Mou, L., Lu, X., 2015. Scene recognition by manifold regularized deep learning architecture. IEEE Transactions on Neural Networks and Learning Systems (TNNLS) 26, 2222–2233.

[43] Zagoruyko, S., Komodakis, N., 2016. Wide residual networks. arXiv preprint arXiv:1605.07146 .

[44] Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., Mech, R., 2015. Minimum barrier salient object detection at 80 fps, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1404–1412.

[45] Zhao, R., Ouyang, W., Li, H., Wang, X., 2015. Saliency detection by multi-context deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1265–1274.

[46] Zhao, R., Oyang, W., Wang, X., 2016. Person re-identification by saliency learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 39, 356–370.

[47] Zhou, F., Bing Kang, S., Cohen, M.F., 2014. Time-mapping using space-time saliency, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3358–3365.