Video Smoke Detection Based on Deep Saliency Network

Gao Xu\textsuperscript{a}, Yongming Zhang\textsuperscript{a}, Qixing Zhang\textsuperscript{a,*}, Gaohua Lin\textsuperscript{a}, Yang Jia\textsuperscript{b}, Jinjun Wang\textsuperscript{a}

\textsuperscript{a}State key Laboratory of Fire Science, University of Science and Technology of China, Hefei 230026, China
\textsuperscript{b}XI’AN University of Posts & Telecommunications, Xian 710121, China

Corresponding author: Qixing Zhang, Email: qixing@ustc.edu.cn, Tel: 86-551-63600770

Abstract

Video smoke detection is a promising fire detection method especially in open or large spaces and outdoor environments. Traditional smoke detection consists of candidate region extraction and classification, but it lacks powerful characterization for smoke. In this paper, we propose a novel method for video smoke detection based on deep saliency network. Visual saliency detection aims to highlight the most important object regions in an image. The pixel-level and object-level salient CNNs are combined to extract the informative smoke saliency map. For the need of application for smoke event detection, an end-to-end framework for salient smoke detection and existence prediction of smoke is proposed. The deep feature map is combined with the saliency map to predict the existence of smoke in image. Initial dataset and augmented dataset are built to measure the performance of frameworks with different design strategies. Qualitative and quantitative analysis at frame-level and pixel-level demonstrates the excellent performance of the ultimate framework.

Keywords

Video smoke detection; deep saliency network; salient map; existence prediction
1. Introduction

Video smoke detection is a promising fire detection method especially in open or large spaces and outdoor environments. The typical traditional video smoke detection methods consists of dynamic texture [1], wavelet based method [2], higher order linear dynamical systems [3], histogram-based smoke segmentation [4], and local extremal region segmentation [5]. As for many other computer vision tasks, recent advances in approaches based on deep networks have resulted in significant performance improvement on the public benchmark datasets. Related work has been focused on the deep network based video smoke detection, such as spatial-temporal ConvNet smoke features [6], [7] based on the combination of motion characteristic and CNN, and deep domain adaptation network [8]. This paper introduces a novel approach for smoke recognition based on saliency detection. To the best of our knowledge, we are the first to investigate smoke detection based on deep saliency network.

Visual saliency detection, which aims to highlight the most important object regions in an image, while semantic segmentation aims to segment objects of certain classes in images. For saliency detection, an object is salient or not is largely depend on its surroundings. It has been widely utilized for many computer vision tasks, such as semantic segmentation, object tracking, and image retrieval, etc.

Overall, saliency detection methods can be divided into traditional methods and CNN based methods. In [9], Jiang decompose the image to multiple segmentation from a coarse one to fine one based on multi-level segmentation, map the region features to saliency score with a random forest regressor and finally fuse the saliency maps across multiple levels of segmentation. Zhu [10] proposes to generate initial maps from color and depth saliency map, center saliency prior and dark channel prior, representatively. Then these initial saliency maps are fused to generate the final saliency map. As reviewed in [11], the numerous salient object detection have been made in recent years.

Since the latest generation of CNNs have substantially outperformed handcrafted approaches in computer vision field, CNN based methods has attracted wide attention for its superior performance. CNN based methods includes region-level methods and pixel-level methods. For the region-level methods, Li [12] proposes to incorporate multiscale CNN features extracted from nested windows with a deep neural network
with multiple fully connect layers. Zhao [13] designs a multi-context deep models on the superpixels of image, including global-context modeling for saliency detection with a superpixel-centered window and local-context modeling with a closer-focused superpixel-centered window. For the pixel-level methods, Cheng [14] propose a new saliency method by introducing short connections to the skip-layer structures within the HED architecture. Li [15] proposes a multi-task FCNN based approach to model the intrinsic semantic properties of salient objects and present a fine-grained super-pixel driven saliency refinement model for the output of the proposed FCNN model. Liu [16] adopt a novel hierarchical recurrent convolutional neural network (HRCNN) to refine saliency maps in details by incorporating local contexts.

More and more deep networks focusing on the combination of multiple branches for information fusion, such as extra features for CNN-based detector in [17], simultaneous detection and segmentation in [18]. The typical application is the combination of handcrafted and deep features, such as [19] integrate handcrafted low-level features with deep contrast features for a more robust feature. Qu [20] design a convolutional neural network (CNN) to automatically learn the interaction between these low-level saliency cues and take advantage of the knowledge obtained in traditional saliency detection. These works lack the analysis on interaction between saliency architecture designs and data. As shown in [21], the performance of the detector is easy to be degraded by the dataset bias. In the experiment section, extensive qualitative and quantitative experimental evaluations on the architecture design and dataset have been executed.

2. Related work

There are many works on the saliency detection with multiple branches. Chen [22] proposes a saliency model built upon two stacked CNNs. The first CNN generates a coarse-level saliency map in the global context. The second CNN integrate superpixel-based local context information in the first CNN to refine the coarse-level saliency map. Li [23] designs a deep network consisting of pixel-level fully convolutional stream and a segment-wise spatial pooling stream. The first stream produces a saliency map in pixel-level based on deeplab [24], and the second stream extracts segment-wise features. Finally, a proposed fully connected CRF model is optionally incorporated to refine the fused result from these two streams. Conditional Markov
random fields (CRFs) are one of the most effective approaches to enforce spatial contiguity in the output label maps. [20] develops a graph Laplacian regularized nonlinear regression scheme for saliency refinement to generate a fine-grained boundary-preserving saliency map. [25] applies the adversarial training to semantic segmentation and optimize the objective function that combines a conventional multi-class cross-entropy loss with an adversarial term. Especially, there are some technologies of spatial coherence refinement for the post-processing approaches, such as CRF [19, 23, 24], clustering [12, 22] using superpixel, graph Laplacian regularized nonlinear regression [15, 20]. [13] outputs the occupation ratio of salient softmax value, while the general saliency network output sigmoid of probability of saliency. All in all, compared to the typical two-stream baseline CRPSD and DCL [23, 26], the differences of our work are as follow: Firstly, the region-level saliency prediction in these methods is mainly used to enhance the edge of the salient object. As the edge of smoke is fuzzy, the pixel-level saliency stream is the basic and the region-level saliency is the partner; Secondly, we have taken the background-only images into consideration for training, while at least one salient object exists in each image of general salient dataset; Thirdly, we have investigated the performance of multiple level salient smoke features.

It is noted that the time consuming is also increasing with the integration of multiple information. We will give the time statistics for measurement. In summary, this paper has the following contributions:

- We investigate the performance differences between state-of-art saliency detection methods for smoke detection.
- We propose an end-to-end framework for salient smoke detection and prediction for smoke existence.
- The integration of multiple-level saliency cues is proposed for smoke detection, and a detailed analysis of the time consuming and performance is provided.
- Qualitative and quantitative evaluations at frame-level and pixel-level demonstrate the excellent performance of the proposed method.

The rest of this paper is organized as follows. In Section 3, we give an overview of the multiple level saliency detection methods for smoke detection. Then in Section 4, the proposed salient smoke detection method is introduced. Section 5 gives the experimental analysis for measurement. And the conclusion and future work
are represented in Section 6.

3. Saliency segmentation for smoke

We investigate the gap of the state-of-art saliency detection methods for smoke detection, including the traditional methods and CNN based methods. The performance of each method is estimated for salient smoke segmentation, and the methods tend to give false alarms at hard negatives, such as fog, lighting, cloud.

3.1 Object-level saliency detection

The object detection framework used for segment output [27] has been researched, there is also work [17] on fusion information from detection model and segmentation model. As introduced in [28], the object proposal algorithms can measure how likely a region contains an object without the need of category information, which can be used to obtain high-level saliency priors. There is a similar case [29] that objectness heatmaps are obtained from the proposals generated by RPN [30]. Typically, the graph-based proposal methods such as selective search segmentation can’t recognize the smoke region well, compared to the pedestrian detection as shown in Fig. 1. In region-level saliency method, the SLIC [31] method is widely used. And there are some slightly modified version of the SLIC algorithm, such as the one used in DCL. The objectness calculations and employment between [29] and [28] are different.

In [29], for each proposal $R_i$ in image, its objectness score $b^i$ is added to all the pixels in the corresponding window. In [28], the pixel distribution namely different distance from candidate box center is considered, the objectness score of each pixel $s_p$:

$$s_p = \left[ \sum_{i=1}^{N} b^i \right] \exp\left\{- \lambda d(p, B_i) \right\}^{1/2}$$

(1)

Where $d(p, B_i)$ is the normalized distance between the pixel $p$ and the center of the candidate box $B_i$ (total number is $N$). $I(p \in B_i)$ indicates whether $p$ is inside the $B_i$. Due to the inconsistent shape of the candidate boxes, the time complexity is $O(n^3)$, which cause that this method is time consuming. In this work, the object proposal method is used to obtain the objectness score of smoke.

Our RPN model is obtained from the Faster-RCNN model [30] trained on the box-level annotation dataset for smoke detection. As shown in Fig. 2, a relatively accurate location information can be provided in the
heatmap. RPN outputs the candidate boxes with confident scores and the objectness score of each pixel is normalized to [0, 255].

Fig. 1. The top row represents the original images, the bottom row represents the initial segmentation results and the heatmap of selective search method.

Fig. 2. From left to right is: original image, heatmap, feature map on conv4_3, feature map on conv5_3, prediction of FCN.

3.2 Region-level segmentation

Region-level segmentation extracts features of regions as context to predict saliency score of each region, typically as superpixels based approach. SLIC algorithm [31] is used to obtain superpixels. As analyzed in [26], if there are too few superpixels, the smoke region will be under-segmented; if there are too many superpixels, the smoke region or background will be over-segmented.

Fig. 3. The superpixels with overlap>0.5 are in white and the little black blocks represent the center of each superpixel.
The top row represents the original images, the middle row represents the segmentation with 100 superpixels and the bottom rows represents the segment with 80 superpixels.

3.3 Traditional saliency smoke detection

In [10], color saliency, depth saliency, center saliency and dark channel prior are fused to obtain the final saliency map. As smoke rarely appear in the center of image and the depth prior is generally not available, we use the dark channel prior (Fig. 4b) with the high dark channel value of smoke and color saliency to recognize smoke (Fig. 4d). Jia [32] propose a saliency-based method for early smoke detection, which calculate the lightness saliency map in HSV and motion saliency map, but experiments showed that it perform not well on the thin smoke and complex scene (Fig. 4c). These traditional methods extract the low-level salient smoke features, which leads to the bad robustness to the complex scenes.

Fig. 4. The top three rows represents the result of dark channel prior saliency (b) and Jia’s method (c). The bottom two rows represents the results of fusion of color saliency and dark channel prior saliency (d).

4. Architecture

There are some work in fusing saliency maps. The saliency maps from both streams are fused at the end through an extra convolutional layer with 1*1 kernels in DCL [23]. EP-MSC [33] obtain the final saliency map by fusing the coarse map via a fully convolutional layer. CRPSD [26] combine the pixel-level saliency and region-level saliency and then generate the final saliency map through three convolutional layers. And it is found that the original image is involved in fusion in some work. The existing methods fusing extra information for salient detection increase detection accuracy and also increase time consumption. Especially
for the two-stream baseline, as the region-level is time consuming, such as the computation of SLIC.

Table 1. Time-consuming statistics. The stage 1 and stage2 are from the superpixels based method [13]. The fusion method runs based on the computed object-level or region-level saliency. Obviously, the SLIC algorithm is too time-consuming.

|          | SLIC   | RPN    | Stage1 | Stage2 | Fusion(matlab) | Fusion(python) |
|----------|--------|--------|--------|--------|----------------|----------------|
|          | 0.4804 | 0.0578 | 0.6994 | 0.5982 | 0.0468         | 0.0166         |

The existing salient object detection assumes that at least one salient object exists in the input image. As smoke appears with low probability, there are many background-only images in the datasets available. This problem is discussed in [34], they adopt the structural SVM framework and integrated salient object detection and existence prediction. As shown in Fig. 6, sometimes the saliency prediction of smoke is some discrete pixels or areas. In this case, it is difficult to judge whether smoke exists in the image. So we propose to integrate the existence prediction branch to the framework. As shown in Fig. 5, the pixel-level CNN is set as basic salient smoke detection using VGG16 [35] convolutional layers. Refer to DHSNet [16], the recurrent convolutional layer is used to incorporate local contexts efficiently in CNN to refine saliency maps. The multiple saliency cues consists of pixel-level saliency, object-level saliency and region-level saliency. As the region-level saliency is much time consuming, it is discarded. The partner CNN branch is used to predict smoke existence by combination of saliency map and feature map.
Fig. 5. The overall framework consists of the master branch (pixel-level CNN) and partner branch for existence prediction. The saliency map is made up of pixel-level, region-level and object-level prediction. The feature map and saliency map are used for existence prediction of smoke.

Fig. 6. The saliency map in the hard samples.

In the framework, the corresponding feature maps and saliency maps are at the multi-scale spatial size \([h^k, w^k]\) and \([h^t, w^t]\) respectively (k, t mean the grade of the layer). For the partner CNN, a convolution operation is applied for the input feature map \(X\) for improving feature representation capability, then a sigmoid activity function is used to control the value of \(X\) in the range of 0 and 1.

\[
U = \frac{1}{1 + e^{-f(X)}}
\]  

(2)

Where \(f(X) = W \times X + b\) is the convolution operation.

The output feature map \(U\) is combined with the resized saliency map through element-wise sum.

Giving an training example \((E, Y_f, Y_p)\) with frame-level label \(Y_f\) and pixel-level label \(Y_p\), the overall loss of the framework consists of the softmax loss on frame-level prediction and cross-entropy loss on pixel-level prediction,

\[
L(Y_f, Y_p) = -\sum_i^{hw} (\alpha \log P(Y_p = 1 \mid p_i) + (1 - \alpha) \log P(Y_p = 0 \mid (1 - p_i))) - \sum_j^{w} Y_f \log f(z_j)
\]

(3)

Where, \(\alpha\) means the ratio of salient pixels in ground truth \(Y_p\), \(p_i\) represents the saliency value of the pixel, and \(z_j\) is the softmax value of the existence prediction.

In general saliency baseline, the region-level stream is used to better model visual contrast between regions and visual saliency along region boundaries. But the contour of smoke is blurred, so whether the region-level saliency prediction of smoke make function need to be confirmed. In our framework, there are three types
of saliency cues: region-level saliency, object-level saliency, and pixel-level saliency. As a reference, we use the dataset with bounding-box level annotation to train the pixel-level salient smoke CNN while the pixels inside the ground-truth box is all the salient pixel.

Firstly, the region-level saliency CNN is trained firstly; with its pre-trained weighted fixed, then the pixel-level saliency CNN is trained, the pixel-level and region-level saliency are fused at the end through an extra convolutional layer with 1*1 kernels. It can be seen in Fig. 7, the region-level saliency make little contribution to the final fusion saliency map, as smoke lacks obvious edge information.

Then, the object proposal CNN is trained firstly; with its pre-trained weighted fixed, then the pixel-level saliency CNN is trained, the pixel-level and object-level saliency are fused at the end through an extra convolutional layer with 1*1 kernels.

Inspired by [34], we jointly train the salient smoke detection and existence prediction for smoke. At the base of the combination of pixel-level and object-level saliency CNN, the partner CNN is built for the existence prediction. As the saliency map offers the probability map of smoke, and the feature map of basic CNN can give the informative representation of the entire image, the combination of them is proposed to extract the feature of high salient candidate to predict the existence of smoke. Different architectures of the overall framework are as follow:

Strategy 1. The pixel-level saliency CNN is trained, with an extra fully connect layer at the end to predict the smoke existence.

Strategy 2. At the base of strategy 2, the original image, the object-level saliency and pixel-level saliency are fused, at the end connecting to 2 convolutional layer and 1 fully connect layer.

Strategy 3. At the base of strategy 2, the high-level saliency map SmRCL4 is combined with the highest level feature map conv4_3, then a spatial pooling layer is used to aggregate information to feature vector and 1 fully connect layer followed.

Strategy 4. At the base of strategy 2, the high-level saliency map SmRCL1 is combined with the sigmoid output of highest level feature map conv4_3, using a spatial pooling layer and eltwise product layer. And a convolutional layer and a fully connect layer followed.
Strategy 5. At the base of strategy 2, a spatial pooling layer and 1 fully connect layer follow the highest level feature map conv4_3.

Strategy 6. At the base of strategy 2, each scale saliency map \((Sm^{RCL1}, Sm^{RCL2}, Sm^{RCL3}, Sm^{RCL4})\) is combined with corresponding feature map \((\text{conv1}_2, \text{conv2}_2, \text{conv3}_3, \text{conv4}_3)\) using element-wise product, then a spatial pooling layer is used to aggregate information to feature vector and 1 fully connect layer followed.

We evaluate the performance using precision-recall (PR) curves as the saliency maps are normalized to \([0, 255]\). The precision and recall are computed by binarizing them with different thresholds and computed them with ground-truth. The commonly used metrics F-measure is computed as:

\[
F_\beta = \frac{(1 + \beta^2) \text{precision} \times \text{recall}}{\beta^2 \text{precision} + \text{recall}}
\]

Where, \(\beta^2 = 0.3\) as the previous work, and the precision and recall are obtained as the saliency map is binarized using the twice the mean saliency value.

5. Experiments

In [21], the local and global color contrast, local gPB boundary strength and object size are used to measure the dataset bias which causes the degradation of performance of salient object detection. In our work, we analyze the following image statistics in order to find the interaction between performance of saliency model and images, including Chi-square distance of RGB histogram, the size of smoke region, thickness of smoke region and dispersion of smoke. In Fig. 7, an initial dataset is built. The images is extracted from the limited collected dataset. The state-of-art deep saliency models are trained and tested at the dataset. It can be seen that the performance of detector is related to the region size and dispersion.
Fig. 7. The top figure represents the image statistic value of each smoke image in test set. And the bottom figure represents the overlap value of saliency prediction and ground-truth annotation in each smoke image of test set.

As far as we know, the data augmentation is rarely mentioned in related work of saliency detection. On the one hand, compared to the saliency dataset rich in scale and diversity, smoke image is extracted from the limited video. On the other hand, we define the smoke detection as salient smoke detection and smoke existence prediction as smoke detection belongs to event detection. To perform a quantitative evaluation, we use both a frame-level (existence prediction) and pixel-level prediction (saliency prediction) for measurement. The test datasets consists of 1399 smoke images and 1401 non-smoke images. The dataset is available at http://smoke.ustc.edu.cn.
Fig. 8. The visual effects of composition using gradient domain cloning. But it is time consuming. In the future work, we will apply some efficient method to overlay images.

The experiments show the evaluation of salient smoke detection with multiple saliency cues. As shown in Fig. 9, the saliency prediction in red is represented in the smoke images (1-4) and non-smoke images (5-7). Fig. 10 shows that in this dataset, the excellent pixel-level salient detections perform well and the results of pixel-level salient smoke detection is close to the fusion saliency detection.

Fig. 9. The top row represents the original images, the second row represents the region-level saliency map (bad performance of region-level saliency for smoke), the third row represents the pixel-level saliency map and the bottom row represents the fusion saliency map of object-level and pixel-level.
Fig. 10. The figures show the F-measure scores (left) and PR curves (right). We represent the pixel-level performances of the state-of-art methods, including fcn with sigmoid output, MC, DHSNet, DSS and the fusion of pixel and region level saliency.

As the dataset is collected from the neighbor frames in video clips, the diversity of images is limited. To enhance the generalization of the saliency model, a new dataset is built using these two measurement:

1. Inspired by Hide and seek [36], the pixel-level ground-truth of smoke region is hided randomly at horizontal and vertical direction. Considering the deformation and media properties of smoke, the occluded smoke images can be used to enhance the diversity of dataset.

2. We collect the hard negative samples such as cloud and fog. Furthermore, the original smoke images are superimposed with the non-smoke images.

Fig. 11. The challenge images in the new dataset. From left to right: the occluded smoke images, the composite smoke images and the hard negative images.

For the new dataset, the joint network for salient smoke detection and existence prediction is trained and
the performance of different design choices is represented as follows. In the new test dataset, there are 2611 images. The 1st ~839th and 1452th~2011th images are smoke images, and 840th ~1451th and 2012th ~2611th are non-smoke images. The dataset is available at http://smoke.ustc.edu.cn. As shown in Fig. 12, the performance of detector is also related to the region size and dispersion.

![Graph](image)

Fig. 12. The top figure represents the image statistic value of each image in the new test dataset, as Fig. 7 shows only the statistic of smoke images. And the bottom figure represents the overlap value of saliency prediction and ground-truth annotation in each image in dataset.
Fig. 13. From top to bottom: the existence prediction precision, the F-measure scores (left) and PR curves (right). We represent the performance of the joint network with different design choices.

Compared to the too well performance of PR curves in Fig. 10 due to limited diversity in the initial dataset,
this PR curves seem to be more reasonable. In can be seen in Fig. 13 that these 5 design choices make little effects on the pixel-level saliency results, but the performances on frame-level prediction of smoke existence is reversed. The strategy 3 achieve the best outcome, as the deep feature map gives more semantic information about smoke region and the top saliency map integrates the fine detail with the global and local information.

6. Conclusion

For video smoke detection, we systematically compare several state-of-art saliency detection methods, including handcraft-feature and CNN based methods. The pixel-level, object-level and region-level salient CNN are combined to extract the informative smoke saliency map. The region-level salient CNN is abandoned due to the little effect and time consuming. For the need of application for smoke event detection, an end-to-end framework for salient smoke detection and existence prediction is proposed. We have built two level datasets for measurement. Qualitative and quantitative evaluations at frame-level (existence prediction) and pixel-level (saliency prediction) demonstrate the excellent performance of the ultimate framework. In the future, the dataset richer in diversity and complex in scene will be created and more efforts on salient smoke detection in video will be put to characterize the smoke saliency in temporal-spatial.

Acknowledgements

This work was supported by the National Key Research and Development Plan under Grant No. 2017YFC0805100, the Anhui Provincial Key Research and Development Plan under Grant No. 1704a0902030, and the Fundamental Research Funds for the Central Universities under Grant No. WK2320000035. The authors gratefully acknowledge all of these supports. The authors specially thank Professor Zhiqiang Zhou for offering the wildfire smoke dataset for us.
Reference

[1] Ye W, Zhao J, Wang S, Wang Y, Zhang D, Yuan Z. Dynamic texture based smoke detection using Surfacelet transform and HMT model. Fire Safety Journal 2015;73:91-101.

[2] Töreyin BU, Dedeoğlu Y, Cetin AE. Wavelet based real-time smoke detection in video. Signal Processing Conference, 2005 13th European: IEEE; 2005. p. 1-4.

[3] Dimitropoulos K, Barmpoutis P, Grammalidis N. Higher Order Linear Dynamical Systems for Smoke Detection in Video Surveillance Applications. IEEE Transactions on Circuits and Systems for Video Technology 2017;27:1143-54.

[4] Krstinić D, Stipaničev D, Jakovčević T. Histogram-based smoke segmentation in forest fire detection system. Information Technology and Control 2009;38.

[5] Zhou Z, Shi Y, Gao Z, Li S. Wildfire smoke detection based on local extremal region segmentation and surveillance. Fire Safety Journal 2016;85:50-8.

[6] Hu Y, Lu X. Real-time video fire smoke detection by utilizing spatial-temporal ConvNet features. Multimedia Tools and Applications 2018:1-19.

[7] Luo Y, Zhao L, Liu P, Huang D. Fire smoke detection algorithm based on motion characteristic and convolutional neural networks. Multimedia Tools and Applications 2017:1-18.

[8] Xu G, Zhang Y, Zhang Q, Lin G, Wang J. Deep domain adaptation based video smoke detection using synthetic smoke images. Fire Safety Journal 2017;93:53-9.

[9] Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S. Salient object detection: A discriminative regional feature integration approach. Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on: IEEE; 2013. p. 2083-90.

[10] Zhu C, Li G, Wang W, Wang R. An Innovative Salient Object Detection Using Center-Dark Channel Prior [C]. IEEE International Conference on Computer Vision Workshop (ICCVW)2017.

[11] Borji A, Cheng M-M, Jiang H, Li J. Salient object detection: A benchmark. IEEE Transactions on Image Processing 2015;24:5706-22.

[12] Li G, Yu Y. Visual saliency based on multiscale deep features. arXiv preprint arXiv:150308663 2015.

[13] Zhao R, Ouyang W, Li H, Wang X. Saliency detection by multi-context deep learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition2015. p. 1265-74.

[14] Hou Q, Cheng M-M, Hu X-W, Borji A, Tu Z, Torr P. Deeply supervised salient object detection with short connections. arXiv preprint arXiv:161104849 2016.

[15] Li X, Zhao L, Wei L, Yang M-H, Wu F, Zhuang Y, et al. Deepsaliency: Multi-task deep neural network model for salient object detection. IEEE Transactions on Image Processing 2016;25:3919-30.

[16] Liu N, Han J. Dhnsnet: Deep hierarchical saliency network for salient object detection. Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on: IEEE; 2016. p. 678-86.

[17] Mao J, Xiao T, Jiang Y, Cao Z. What Can Help Pedestrian Detection? arXiv preprint arXiv:170502757 2017.

[18] Hariharan B, Arbeláez P, Girshick R, Malik J. Simultaneous detection and segmentation. European Conference on Computer Vision: Springer; 2014. p. 297-312.

[19] Li G, Yu Y. Visual saliency detection based on multiscale deep CNN features. IEEE Transactions on Image Processing 2016;25:5012-24.

[20] Qu L, He S, Zhang J, Tian J, Tang Y, Yang Q. RGBD salient object detection via deep fusion. IEEE Transactions on Image Processing 2017;26:2274-85.

[21] Li Y, Hou X, Koch C, Rehg JM, Yuille AL. The secrets of salient object segmentation. Georgia Institute of Technology; 2014.

[22] Chen T, Lin L, Liu L, Luo X, Li X. Disc: Deep image saliency computing via progressive representation learning. IEEE transactions on neural networks and learning systems 2016;27:1135-49.

[23] Li G, Yu Y. Deep contrast learning for salient object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition2016. p. 478-87.

[24] Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv preprint arXiv:160600915 2016.

[25] Luc P, Couprie C, Chintala S, Verbeek J. Semantic segmentation using adversarial networks. arXiv preprint arXiv:161108408 2016.
[26] Tang Y, Wu X. Saliency detection via combining region-level and pixel-level predictions with cnns. European Conference on Computer Vision: Springer; 2016. p. 809-25.

[27] Dai J, He K, Sun J. Instance-aware semantic segmentation via multi-task network cascades. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition2016. p. 3150-8.

[28] Jia Y, Han M. Category-independent object-level saliency detection. Computer Vision (ICCV), 2013 IEEE International Conference on: IEEE; 2013. p. 1761-8.

[29] Ge W, Yang S, Yu Y. Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning. arXiv preprint arXiv:180209129 2018.

[30] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems2015. p. 91-9.

[31] Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence 2012;34:2274-82.

[32] Jia Y, Yuan J, Wang J, Fang J, Zhang Q, Zhang Y. A saliency-based method for early smoke detection in video sequences. Fire technology 2016;52:1271-92.

[33] Wang X, Ma H, Chen X, You S. Edge Preserving and Multi-Scale Contextual Neural Network for Salient Object Detection. IEEE Transactions on Image Processing 2018;27:121-34.

[34] Jiang H, Cheng M-M, Li S-J, Borji A, Wang J. Joint Salient Object Detection and Existence Prediction. Front Comput Sci 2017.

[35] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014.

[36] Singh KK, Lee YJ. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization. arXiv preprint arXiv:170404232 2017.
