DEEP SALIENCY MAP ESTIMATION OF HAND-CRAFTED FEATURES

Guoqing Jin[⋆†] *Shiwei Shen*[⋆†] *Dongming Zhang*[⋆†] *Wenjing Duan*[‡] *Yongdong Zhang* [⋆†]

[⋆] Inst. of Computing Teth.Chinese Academy of Sciences, China
[†]University of Chinese Academy of Sciences, China
[‡]Beijing University of Posts and Telecommunications, China

ABSTRACT

Saliency detection that utilizes deep convolutional neural networks to obtain high level features from original images has achieved considerable progress during the past years. However, few methods consider learning saliency cues from hand-crafted features. In this paper, we demonstrate that deep learning can produce good enough saliency detection results using only hand-crafted features. We propose a novel multi-context deep learning saliency detection algorithm, where only hand-crafted features are taken into account and modeled in a unified deep learning framework. Extensive experiments on benchmark datasets indicate significant and consistent improvements over the representative deep learning framework based saliency detection methods.

Index Terms— Saliency detection, Deep learning, Hand-crafted features

1. INTRODUCTION

Saliency detection aims to select important visual regions in accordance with human attention. The output of an algorithm usually is a map that predicts not only scene locations where an observer may gaze, but also the probability of that pixel belonging to the salient object[1]. It has attracted a great deal of interest in computer vision community because of its wide applications to image editing techniques[2, 3], image retrieval[4], object detection and recognition[5].

Traditional saliency detection methods mainly rely on three aspects: meaningful feature represention, potential saliency cues and optimal integration strategy. Previous works usually model saliency via constrast of hand-crafted features(e.g., color, texture and edge orientation[6, 7]) and human designed meachanisms(e.g., spectral residual[8] and low rank matrix[9]). Inspired that humans tend to gaze at the center of images, Itti et al.[6] proposals one of the earliest saliency models based on center-surround mechanisms. Other commonly used prior knowledge includes background prior[10, 11] and objectness prior[1]. Various integration methods have explored to further improve saliency detection results. Mai et al.[12] uses a Conditional Random Field (CRF) framework to model the contribution from individual saliency map. A graph-based manifold ranking model[11] was contructed to estimate the saliency by ranking the similarity of the image elements (pixels or regions).

The success of deep learning in object classification[13] and recongnition[14] brought to a revolution in saliency detection. MDF(Multiscale Deep Features)[15], MCDL(Multi-Context Deep Learning)[16] and LEGS(Local Estimation and Global Search)[17] have achieved significant improvements on public benchmarks compared with traditional methods because of deep convolutional neural networks obtain more robust features than hand-crafted ones for salient object detection. However, the convolution and pooling operations of CNNs would ”blur” the object boundaries, thus saliency maps typically are too coarse to give fine details. To tackle this problem, ELD-HF(Encoded Low level Distance Map and High Level Features)[18] directly uses deep learning as an encoder architecture to encode low level distance map of superpixels.

All these deep methods have focused on designing appropriate architectures for extracting saliency maps from the original images, and few methods consider learning saliency cues from hand-crafted features. Whether deep learning framework works likes human-being could perceive saliency regions from hand-crafted visual images? Whether hand-crafted features are capable of achieving considerable results in the framework of deep learning? Whether deep learning and human knowledge could be combinated effectively to tackle the cognitive problem?

Based on the above motivations, we propose a new multi-context deep learning algorithm where only hand-crafted features are concatenated automatically through a supervised learning scheme. The hand-crafted features is learned by integrating both posterior fusion estimation and anterior fusion estimation. In the posterior fusion estimation stage, each hand-crafted map is trained seperately to predict the saliency of high level object concepts. In the anterior fusion estimation stage, we use an encoder architecture to encode various hand-crafted features to determine the saliency values that is sensitive to the object boundaries. The posterior fusion estimation and anterior fusion estimation are integrated into the deep learning framework for saliency detection, and are jointly optimized. Quantitative and qualitative experiments

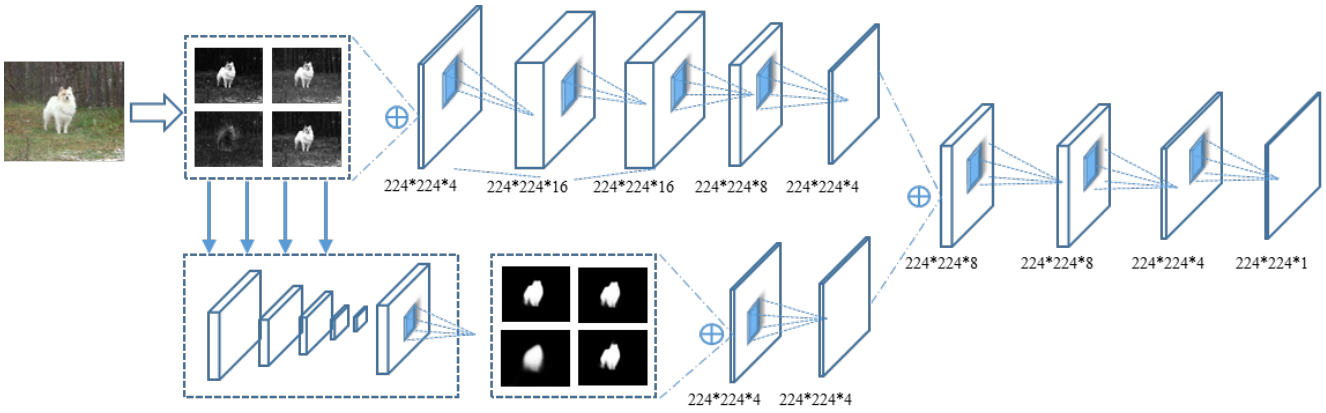


Fig. 1. Overall pipeline of our method. We use four hand-crafted feature maps as the input of the network and learn the saliency values in a dual-estimation process. The upper estimation is fed with fused maps to learn values with precise boundaries. The lower estimation is fed with feature map seperately to learn the object conception saliency. The two features are concated and integrated into a jointly optimized network for final saliency detection.

on four benchmark datasets ASD, PASCALS, ECSSD and HKU-LS demonstrate that our proposal performs favorably against representative CNN-based saliency detection algorithms.

2. PROPOSED SCHEMES

The overall pipeline of our algorithm is illustrated in Figure 1. Our multi-context deep learning network consists of two complementary components, posterior fusion estimation and anterior fusion estimation. The posterior fusion estimation is responsible for saliency cues with precise boundaries, and the anterior fusion estimaion is to learn high level semantic features for detecting salient objects. The saliency maps from both streams are fused at the end through extra fully convolutional layers to produce the final saliency map. The weights in this fusion layers are learned during training.

2.1. Deep Estimation

Our algorithm utilizes four hand-crafted saliency information cues: local constrast(LC)[7], color constrast(CC)[7], spatial variance(SV)[19] and central variance(CV)[19] as the low level features for image saliency. The important principle of feature selection is easy to implement, meanwhile experiments demonstrate that other hand-craft features rarely improve the performance of our algorithm. Both global and local cues were employed since they can provide the saliency info as a supplement for each other. To generate a precise saliency mask, we design a deep estimation architecture consisting of two components, posterior fusion estimation and anterior fusion estimation.

The upper branch of our saliency detection pipeline is a deep encoder architecture of hand-crafted feature. Because the hand-crafted feature maps are concated as the input of the pipeline, we call this operation as anterior fusion estimation. After the global and local cues computation, the initial hand-crafted low level features are concated to $224 \times 224 \times 4$.

As illustrated in Figure 1, the initial feature map is then learned using the multiple 3×3 convolutional and ReLU layers. Regions exceeding image boundaries are padded with mean pixel value of the training dataset. In our implementation, the size of the encoded low level feature map is $224 \times 224 \times 4$, which is consistent with posterior fusion estimation feature dimension.

The lower branch of our algorithm is a posterior fusion estimation process. We utilize FCN-8s model pretrained by VGG to extract the high level features from each hand-crafted feature map seperately. The input maps were fixed to 224×224 of the pretrained model. We discard the final segmental layer and append a 1×1 convolution with a single channel to predict saliency score for each hand-crafted feature map(including background and saliency regions). We further stack these four feature maps together. The stacked feature maps (4 channels) are fed into a final convolutional layer with a 1×1 kernel and four output channels, which are the inferred saliency map of the posterior fusion estimation.

The output of the two branches are concatenated to be a $224 \times 224 \times 8$ feature map. Here, we use the operation concate for concatenating channels. Although the stacked eight output maps have the same size, they are generated using different sizes receptive fields and fusion method.

As shown in Figure 1, we append three extra convolutional layers to the stacked layer. The first extra layer has 3×3 kernels and 8 channels, the second extra layer has 1×1 kernels and 4 channels, and the third extra layer(output feature map) has a 1×1 kernel and a single channel. All of the extra layers are followed by ReLU layers to find the best non-

	ASD	PASCAL-S	ECSSD	HKU-LS
HF_a	0.842	0.524	0.571	0.584
HF_p	0.932	0.802	0.856	0.841
HF	0.941	**0.815**	**0.860**	**0.847**
RGB	**0.951**	0.757	0.826	0.814

Table 1. The maximum F-measure scores of the controlled experiments. Using hand-crafted features shows better performance than using RGB images.

linear representation of the fusion features and generate the final saliency map. The key characteristic of our integrated network is that we utilize fully convolutional layers without any feature map size decrease. Therefore, our method is able to avoid the segement edge blur brought by deconvolution.

A binary classifier network layer is trained to separate background and saliency by minimizing the cross entropy loss for softmax

$$L = -\sum_{j=0}^{1} \mathbf{1}_{(y=j)} \log(\frac{e^{z_j}}{e^{z_0} + e^{z_1}}) \quad (1)$$

where z_0 and z_1 are the non-salient score and slaient score of each label of the training image.

2.2. Training

We utilize 9000 images sampled from MSRA10K dataset after excluding images in ASD dataset to train our network, and each time only one image is loaded. The model is trained until its training data loss converges without any validation.

We adopt two steps to train our network. In the first step, we train the posterior fusion estimation stream. The model is initialized with a pre-trained FCN-8s model from $conv1\_2$ layer to $fc7$ layer and fine-tuned end-to-end for the saliency detection task. In the next step, the anterior fusion estimation layers is trained end-to-end with the posterior fusion estimation stream fixed by setting the learning rate equal to zero. The weights of anterior fusion estimation layer were initialized by the *xavier*. We set the base learning rate equal to $1e$-10 and adopt stochastic gradient descent method with momentum 0.95, weight decay 0.0005.

2.3. Analysis of the Efficiency

It is much common for deep neural networks to generate feature maps from original images. Compared with original images, hand-crafted feature maps have better aggregation, which are more likely to help CNN to generate the fine-grained dense saliency mask.

To demonstrate the effects of the deep estimation of the hand-crafted features in our algorithm, we performed multiple controlled experiments. We conducted the experiments

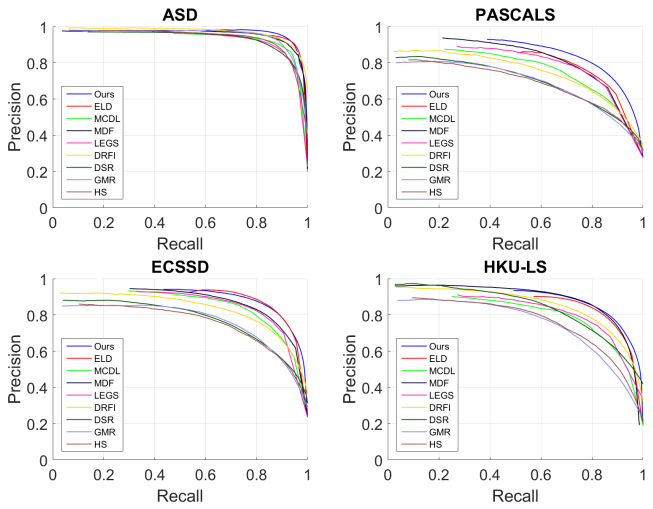


Fig. 2. Quantitative PR-curve evalution of 9 approaches on 4 datasets. Our method performs effectively against others.

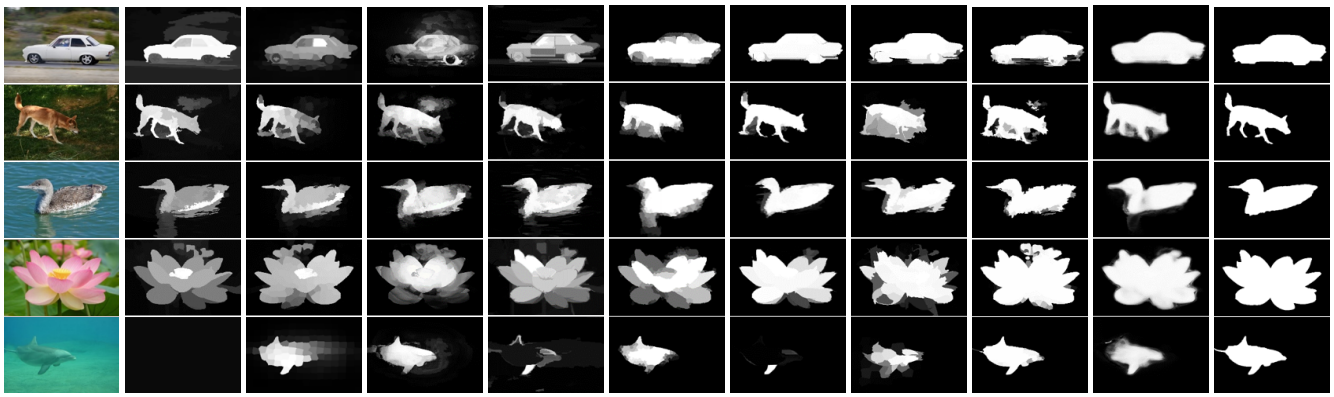
	ASD	PASCALS	ECSSD	HKU-LS
Ours	**0.024**	**0.090**	**0.084**	**0.097**
ELD	**0.025**	**0.101**	**0.080**	**0.076**
MCDL	0.037	0.130	0.097	0.098
MDF	0.053	0.108	0.109	0.109
LEGS	0.064	0.115	0.102	0.106
DRFI	0.085	0.196	0.166	0.115
DSR	0.080	0.205	0.173	0.118
GMR	0.075	0.217	0.189	0.132
HS	0.111	0.262	0.228	0.157

Table 2. The MAE of salient region detection algorithms on four popular datasets. The best two results are shown in red and blue respectively.

using four different settings: The HF_a uses anterior fusion estimation of hand-crafted features. The HF_p utilizes posterior fusion estimation of hand-crafted features. The RGB setting utilizes both the anterior fusion estimation of hand-crafted features and the high level feature map of RGB images from the FCN-8s model. The HF uses the proposed architecture of this paper. The results of the controlled experiments are shown in Table 1. The model utilizing hand-crafted features exhibits better performance than original RGB images.

3. EXPERIMENT

In this section, we detail the implementation of our experiments. We evaluate the proposed method on benchmark datasets and compare with the classic or representative methods.



(a) Image (b) HS (c) GMR (d) DSR (e) DRFI (f) LEGS (g) MDF (h) MCDL (i) ELD (j) Ours (k) GT

Fig. 3. Qualitative comparison of the results of our approach(Ours) with gournd truth(GT) and several other algorithms. The results of our algorithm are the most consistent with the ground truth.

	ASD	PASCALS	ECSSD	HKU-LS
Ours	**0.941**	**0.815**	**0.860**	**0.847**
ELD	**0.930**	**0.784**	**0.868**	0.839
MCDL	0.904	0.743	0.822	0.780
MDF	0.921	0.782	0.832	**0.849**[1]
LEGS	0.895	0.772	0.827	0.788
DRFI	0.925	0.716	0.790	0.806
DSR	0.886	0.675	0.742	0.785
GMR	0.921	0.677	0.750	0.733
HS	0.904	0.669	0.734	0.741

Table 3. The maximum F-measure scores of salient region detection algorithms on four popular datasets. The best two results are shown in red and blue respectively.

The evaluation is conducted on four typical datasets: ASD[20], PASCALS[1], ECSSD[21] and HKU-LS[15]. It is important to note that all datasets are challenging and widely used. We compare our algorithm with saliency detection algorithms including HS[21], GMR[11], DSR[22], DRFI[23], LEGS[17], MDF[15], MCDL[16] and ELD[18], which are the representative algorithms. HS, GMR, DSR and DRFI utilize low level context while LEGS, MDF, MCDL and ELD use deep learning based high level features. For fair comparison, we use either the saliency maps provided by the authors or the implementation code with default parameters. We evaluate the performance of our algorithm using precision-recall curve, F-measure methodologies and Mean Absolute Error(MAE)[24].

Figure 2 presentes the comparisons on Precision-Recall graph. We observe that our algorithm achieves the better performance than the previous works in terms of precision-recall curve. Maximum F-measure scores and MAE values are also described in Table 3 and Table 2. Our algorithm shows the highest maximum F-measure score and the lowest MAE on most of the datasets.

A visual comparison of various methods are shown in Figure 3. As it can be seen, our algorithm produces more accurate saliency maps in all sorts of difficult scenarios, especially good performance on images with low-contrast objects. We explore that our method inherently highlights the salient region and preserve explicit object boundary than other methods. The robust performance of our method can be contribute to the use of the integration of the posterior fusion estimation and anterior fusion estimation.

[1] MDF obtained highest maximum F-measure score in HKU-LS because they used 3000 images in HKU-LS to train the model.

4. CONCLUSION

We present a new multi-context learning network for deep saliency map estimation by only using hand-crafted features. We utilize four hand-crafted feature maps as the input of the network and learn the saliency values in a dual-estimation process that model precise boundaries and high level semantic salient objects separately. Experimental results validate the effectiveness of the proposed model which uses hand-crafted features as image saliency cues.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Plan of China (2016YFB0801203, 2016YFB0801200), in part by the National Natural Science Foundation of China (No.61672495, No.61379084, No.61402440).

6. REFERENCES

[1] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE CVPR*, 2014, pp. 280–287.

[2] Fred Stentiford, "Attention based auto image cropping," in *Workshop on Computational Attention and Applications, ICVS*. Citeseer, 2007, vol. 1.

[3] Luca Marchesotti, Claudio Cifarelli, and Gabriela Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *2009 IEEE 12th ICCV*. IEEE, 2009, pp. 2232–2239.

[4] Yue Gao, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 363–376, 2013.

[5] Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona, "Is bottom-up attention useful for object recognition?," in *CVPR 2004.* IEEE, 2004, vol. 2, pp. II–37.

[6] Laurent Itti, Christof Koch, Ernst Niebur, et al., "A model of saliency-based visual attention for rapid scene analysis," *IEEE TPAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[7] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu, "Global contrast based salient region detection," *IEEE TPAMI*, vol. 37, no. 3, pp. 569–582, 2015.

[8] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *2007 IEEE CVPR*. IEEE, 2007, pp. 1–8.

[9] Xiaohui Shen and Ying Wu, "A unified approach to salient object detection via low rank matrix recovery," in *2012 IEEE CVPR*. IEEE, 2012, pp. 853–860.

[10] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun, "Geodesic saliency using background priors," in *European Conference on Computer Vision*. Springer, 2012, pp. 29–42.

[11] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE CVPR*, 2013, pp. 3166–3173.

[12] Long Mai, Yuzhen Niu, and Feng Liu, "Saliency aggregation: A data-driven approach," in *Proceedings of the IEEE CVPR*, 2013, pp. 1131–1138.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[15] Guanbin Li and Yizhou Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE CVPR*, 2015, pp. 5455–5463.

[16] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE CVPR*, 2015, pp. 1265–1274.

[17] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE CVPR*, 2015, pp. 3183–3192.

[18] Gayoung Lee, Yu-Wing Tai, and Junmo Kim, "Deep saliency with encoded low level distance map and high level features," *arXiv preprint arXiv:1604.05495*, 2016.

[19] Tiancai Ye, Dongming Zhang, Ke Gao, Guoqing Jin, Yongdong Zhang, and Qingsheng Yuan, "Salient region detection: Integrate both global and local cues," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2014, pp. 1–6.

[20] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk, "Frequency-tuned salient region detection," in *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*. IEEE, 2009, pp. 1597–1604.

[21] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE CVPR*, 2013, pp. 1155–1162.

[22] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang, "Saliency detection via dense and sparse reconstruction," in *Proceedings of the IEEE CVPR*, 2013, pp. 2976–2983.

[23] Huaizu Jiang, Jingdong Wang, Zejian Yuan, and Yang Wu, "Salient object detection: A discriminative regional feature integration approach," *International Journal of Computer Vision*, vol. 9, no. 4, pp. 1–18, 2014.

[24] Ali Borji, Dicky N. Sihite, and Laurent Itti, *Salient Object Detection: A Benchmark*, Springer Berlin Heidelberg, 2012.