24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Saliency Detection with Deformable Convolution and Feature Attention

Zhe Zhang[1,2] and **Junhui Ma**[1,2] and **Panpan Xu**[1,2] and **Wencheng Wang**[1,2,3]

Abstract. Recently, with the development of Convolutional Neural Networks (CNNs), deep learning-based saliency detection methods have advanced significantly. Most of the existing deep learning-based methods attempt to extract semantic context information to yield a saliency map. However, it is difficult to capture irregular context features by using a standard convolution because such features are often unevenly distributed in complex scenes. To address this problem, this paper proposes a novel saliency detection model named DCFA, which is implemented using two important modules. First, we design a Deformable Feature Extraction Module (DFEM) to focus on the unevenly distributed context features in both low-level details and high-level semantic information. Second, a Channel and Spatial Attention Module (CSAM) is devised to assign the adaptive weights of the features in the space and channel domains. The experimental results show that the proposed model can achieve the state-of-the-art performance on six widely used saliency detection benchmarks. Furthermore, our proposed network is end-to-end and runs at a speed of 20 fps on a single GPU.

1 Introduction

Saliency detection aims to locate the attractive and interesting regions in images, which plays an important role in many applications, such as person re-identification [3], visual tracking [13] and image segmentation [10]. Considerable research has been performed in recent years, leading to significant development in saliency detection. Conventional approaches [5, 36] usually design hand-crafted low-level features and make heuristic hypothesizes, which often fail in obtaining satisfactory results in complex scenes. Recently, deep learning-based methods [14, 25, 27, 41] have made significant improvements in saliency detection because convolutional neural networks (CNNs) can learn high-level semantic features. Hence, semantic context features are crucial for saliency detection under complex scenes. Hou et al. [14] combined the low-level and high-level features using short connections to predict the saliency maps. Zhao et al. [41] used dilation convolution with different rates to extract multi-scale features to yield more accurate saliency maps. However, the semantic context features are often unevenly distributed in images, and thus these methods cannot be used to extract the features accurately because of the limitation of the standard convolution in the CNNs.

Deformable convolution [7] modifies the fixed shape of the standard convolution by introducing a set of offsets to shift the location of the input features, which enables it to adaptively extract context features. However, this paper mainly concentrates on creating the deformable convolution layer through an extra offset layer, it does not discuss how to utilize the deformable convolution layer properly in specific vision tasks. In this paper, we propose a novel saliency detection model to better extract important features with deformable convolution and feature attention, named DCFA. The DCFA involves two important modules: (1) The Deformable Feature Extraction Module (DFEM) can detect the context features that are irregularly shaped owing to the deformable convolution. These context features extracted using the DFEM can overcome the limitation of standard convolution, which can significantly improve the saliency detection performance. (2) The Channel and Spatial Attention Module (CSAM) can learn adaptive weights of different features. Specifically, spatial attention focuses on the most salient regions in the space domain and it can filter out background noises, and channel attention can select more semantic meaningful features in the channel domain. The design of the proposed modules is motivated by the following two aspects:

First, salient objects generally have different scales and shapes. The recent deep saliency detection models mainly focus on combining the outputs from the intermediate network layers. Thus, although such simple integrations may help extract multi-scale features, the unevenly distributed context features cannot be well detected because of the limitation of standard convolution. Unlike the existing approaches, we propose a novel Deformable Feature Extraction Module (DFEM) to detect the irregularly distributed context features. Specifically, we adopt the deformable convolution [7] in different layers to locate the unevenly distributed salient features, which can significantly improve the quality of saliency prediction.

Second, the different features in CNNs usually exert different influences. Low-level features generally have structural details but also contain noises, whereas high-level features often carry rich semantic information along with unimportant ones. These noises or unimportant features will prevent the generation of precise saliency maps. However, many existing methods integrate such features without any distinction, thereby leading to an inaccurate prediction. Inspired by [4], which uses the channel and spatial attention to improve the results of image caption, we design a Channel and Spatial Attention Module (CSAM) to extract the most meaningful features adaptively in different layers. The CSAM can highlight the crucial features and suppress the unnecessary ones, which is essential for our model.

Our main contributions can be summarized as follows:

- We develop a Deformable Feature Extraction Module (DFEM), which can capture the unevenly distributed context features to improve the saliency prediction.
- We design a Channel and Spatial Attention Module (CSAM) to select the most salient features and suppress the noises in the space and channel domains.

[1] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, China
[3] Corresponding author: whn@ios.ac.cn (Wencheng Wang)

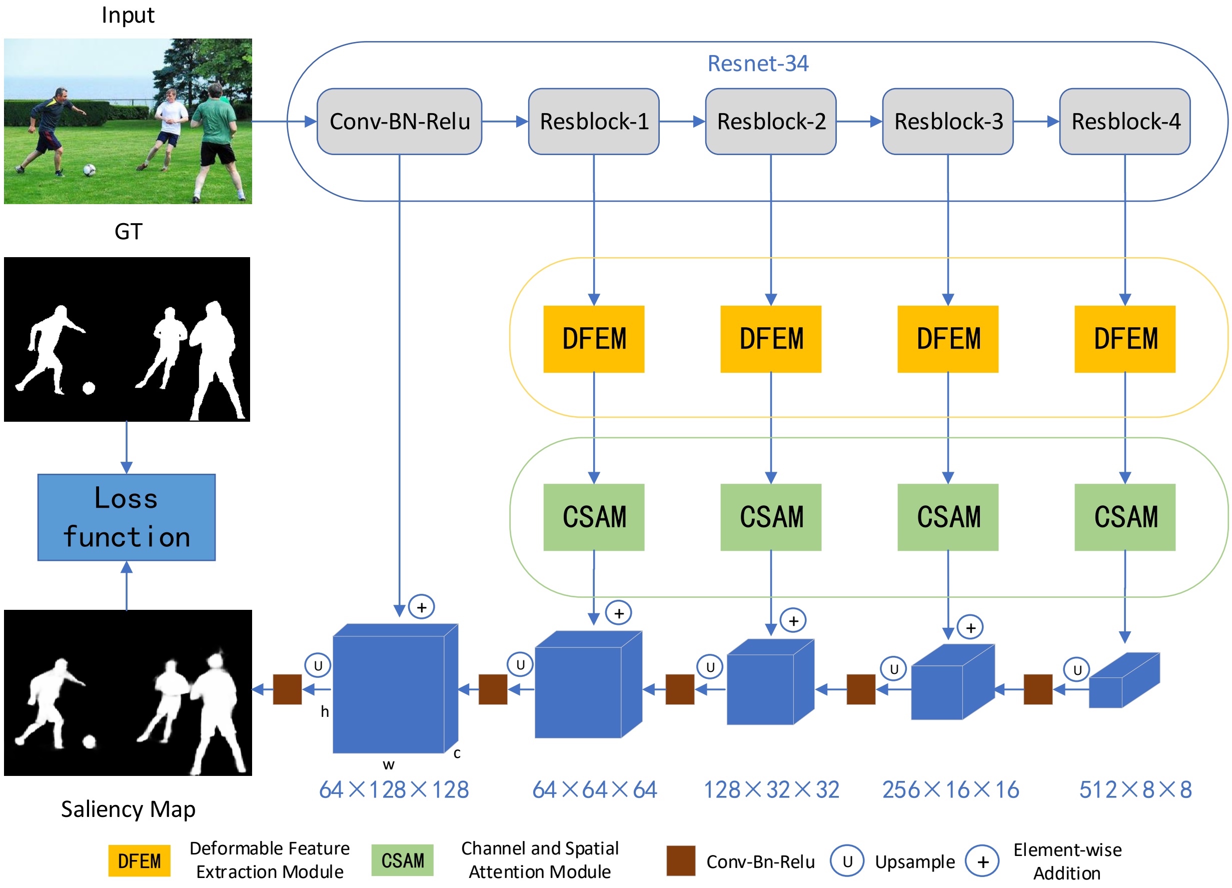


Figure 1: Overall architecture of the proposed DCFA.

- We compare the proposed DCFA with 13 state-of-the-art approaches on six widely used datasets. The experimental results demonstrate that the proposed method can achieve the state-of-the-art performance under different evaluation metrics.

The remaining paper is organized as follows. Section 2 provides a review of the related work. Section 3 describes the architecture of the proposed method. Section 4 shows the experimental results. Section 5 presents the conclusion.

2 Related Work

2.1 Saliency Detection

Early saliency detection methods [5, 35, 36] are mostly based on the low-level features such as color, texture and heuristic priors, but the hand-crafted features and simple priors make it difficult to capture the high-level semantic information. For example, Wei et al. [30] proposed a boundary prior to measure the saliency of each superpixel via the geodesic distance of the boundary. However, such methods often fail when the saliency region is at the boundary of the image. To solve the boundary prior failure problem, Zhu et al. [42] proposed a boundary connectivity prior approach, in which a higher salient value is assigned to the region with fewer boundary connections.

In recent years, due to the success of CNNs in computer vision, deep learning-based methods have been widely used for saliency detection. These models mainly employ the semantic information to obtain the global saliency information. Cheng et al. [14] proposed a novel saliency method in which short connections are introduced to the skip-layer structures within the HED [33] architecture. In [14], instead of connecting the loss layers directly to the last layer of each stage, a series of short connections are introduced between the shallower and deeper side-output layers, and the activation of each side output layer is employed to highlight the entire salient object and precisely positions its boundaries. Zhang et al. [38] developed a generic aggregating multi-level convolutional feature framework for saliency detection. Luo et al. [21] proposed an approach that further improves the edge accuracy by adding a boundary loss term to the typical cross-entropy loss. Deng et al. [9] proposed a new recursive residual-refinement network equipped with a residual-refinement block to more accurately detect the salient regions of the input images. However, these methods cannot extract the unevenly distributed features since standard convolution can only capture the features in regular domains.

2.2 Attention Mechanisms

Attention mechanisms have been successfully applied in various tasks such as machine translation [11], pose estimation [6], and visual question answering [34,37]. Bahdanau et al. [2] developed an attention model with differentiable soft alignments for machine translation. In recent years, attention models have been applied to several vision tasks. Sermanet et al. [24] determined the participation region via a recurrent attention model for fine-grained classification. Chu et al. [6] proposed the incorporation of CNNs with a multi-context attention mechanism into an end-to-end framework for human pose estimation. These works demonstrated that attention mechanisms can facilitate saliency detection tasks by attending to information context.

Zhang et al. [40] proposed an attention-guided network that selectively integrates multiple levels of context information via the channel and spatial attention model. Wang et al. [29] devised an essential pyramid attention structure for salient object detection, which enables the network to focus more on salient regions while exploiting the multi-scale saliency information. Since attention mechanisms have a great ability to effectively select features, it is suitable for saliency detection. Inspired by [4], we adopt the channel and spatial attention to choose the most salient features in both the channel and space domains.

3 Proposed Model

In this paper, we propose a novel saliency detection model named DCFA, which includes a Deformable Feature Extraction Module (DFEM) and a Channel and Spatial Attention Module (CSAM). The DFEM focuses on capturing the unevenly distributed context features. The CSAM pays attention to assign larger weights to the most salient features and weaken the weights of the unimportant ones in the channel and space domains. We use the pre-trained Resnet-34 [12] as our base feature extraction network. The overall architecture of DCFA is shown in Figure 1.

3.1 Deformable Feature Extraction Module

The context feature is important for saliency detection. However, salient objects usually vary considerably in terms of the scale and shape, which is a challenging problem in saliency detection. Previous deep learning-based models try to obtain different features by stacking multiple standard convolutional layers, which is inefficient to handle these complicated scenes, especially the unevenly distributed salient objects. As shown in [7], the deformable convolution can capture the irregular features, so we design the Deformable Feature Extraction Module (DFEM) to capture the scale and shape variation of features.

Deformable convolution, which was first proposed in [7], can augment the spatial sampling locations in the feature layers with additional offsets and learn the offsets from the target tasks. However, it mainly focuses on how to build the deformable convolution layer through an extra offset layer. The authors do not discuss how to create the deformable convolution layer in specific vision tasks. For example, the authors simply used deformable convolution layer on high-level feature layers while ignoring the low-level features. In saliency detection, we found that the deformable convolution layer can be applied in both low-level and high-level feature layers, by which it can produce more convincing saliency predictions.

Table 1: Deformable convolution settings in DFEM.

Layers	Kernel settings
Resblock-1	{de-conv 64x7x7, de-conv 64x3x3}
Resblock-2	{de-conv 128x5x5, de-conv 128x3x3}
Resblock-3	{de-conv 256x3x3, de-conv 256x3x3}
Resblock-4	{de-conv 512x3x3, de-conv 512x3x3}

Table 2: Experiment results using different numbers of deformable convolution (de-conv) layers. A higher F_β^{max} and lower MAE corresponding to better results.

Number of de-conv layers	Training time/hour	F_β^{max}	MAE
1	9	0.936	0.042
2	12	0.940	0.038
3	18	0.941	0.039

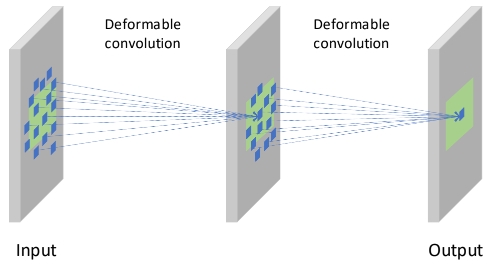


Figure 2: Illustration of the deformable convolutional layer, taking the DFEM after Resblock-3 as an example. It can be seen that the deformable convolution layer can extract the irregular features.

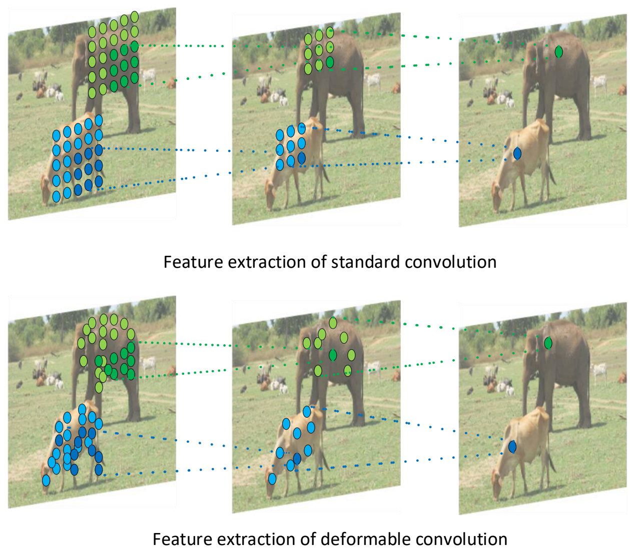


Figure 3: Feature extraction through standard and deformable convolutions. It turns out that deformable convolution can better detect the unevenly distributed features of animals, such as the legs of the sheep.

The deformable convolution layer is shown in Figure 2. We only use the features from Resblock-1, 2, 3, and 4 since the features produced by the first convolution layer are excessively rough. Because the sizes of the features of each block are different, kernels with different sizes can be used to extract the multi-scale context features.

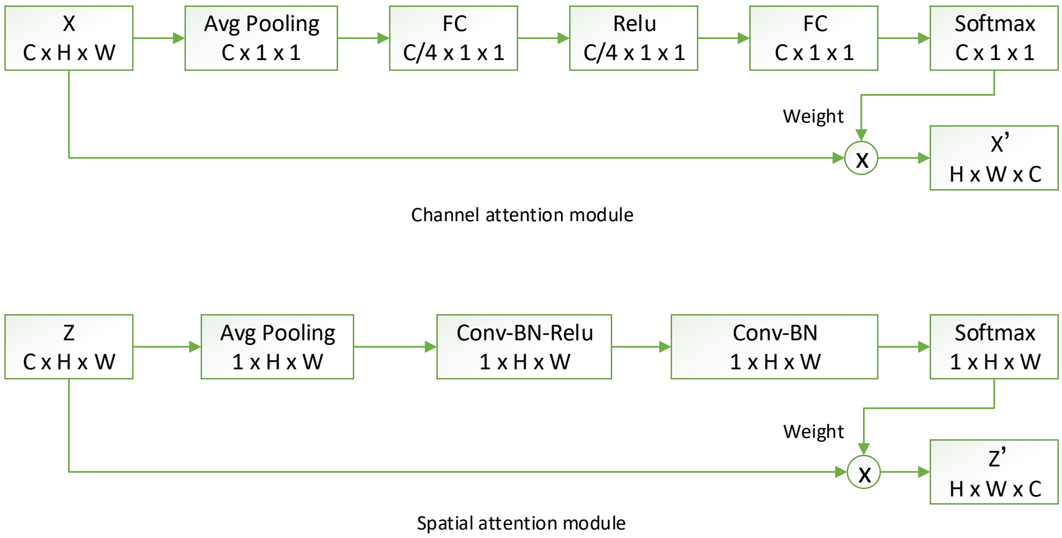


Figure 4: Channel and spatial attention module. Note that in the channel and spatial attention modules, the Avg Pooling is along the h, w and channel directions, respectively.

In this paper, we use the 7×7, 5×5 and 3×3 deformable convolution kernels in DFEM, and their details are presented in Table 1. Note that the deformable convolution is computational in the training period, and thus using two or three deformable convolution layers can yield better results than using only one. Besides, the results obtained by using two or three deformable convolution layers are closed, the reason is that the receptive field is sufficiently large to extract the irregular features when using two or three deformable convolution layers. However, the use of three deformable convolution layers takes a longer training period than using two layers, as indicated by the experiment results in Table 2. Hence, we use two deformable convolution layers after each Resblock in the final experiments.

Figure 3 shows the different effects of the standard and deformable convolutions. It can be seen that the standard convolution is fixed for all aspects of the feature, while the deformable convolution is adaptively adjusted according to the objects' scale and shape, which is helpful to generate better saliency maps.

3.2 Channel and Spatial Attention Module

Given an image, it is obvious that the extracted features have different influences on the final saliency map. The channel attention focuses on what the salient object is while spatial attention pays attention to where the salient object is. Therefore, we need to find the inter-channel and inter-spatial relationships to locate important features. The details of our channel and spatial attention module are shown in Figure 4.

3.2.1 Channel Attention Module

The different channels of the features in CNNs generate different responses for different semantics. The channels contain various structural details for low-level features and different semantics for high-level features. Thus, it is necessary to focus on the important features and weaken the unimportant ones. We add the channel attention module (CAM) after the DFEM to assign adaptive weights to the features. The CAM assigns larger weights to the channels that show a high response to salient objects, and it can be represented as follows:

$$CAM = Softmax(fc_2(\sigma(fc_1(AvgPool(X), W_1)), W_2))$$

$$X' = CAM \odot X$$

where $X \in \mathbb{R}^{C \times H \times W}$ is a feature map, CAM is the output of the channel attention module, AvgPool is the average pooling along the H, W direction, fc_1 and fc_2 are the fully connected layers that capture the channel dependencies, W_1 and W_2 are the respective weights of fc_1 and fc_2, σ is the $Relu$ non-linear activation function, and the $Softmax$ function is used to enhance the most salient channel and weaken the non-salient channel. Finally, the weighted feature X' is calculated by performing an element multiplication between CAM and the original feature X.

3.2.2 Spatial Attention Module

Natural images usually contain a wealth of details of foreground and complex background. Low-level features contain several details, while high-level features may include background noises that may lead to inferior results. In saliency detection, the objective is to identify detailed boundaries between the salient objects and the background without other textures that can distract human attention. Therefore, instead of considering all the spatial positions equally, we adopt the spatial attention module after the DFEM to focus more on the salient regions, which helps to generate effective features for saliency prediction. The SAM can be described as follows:

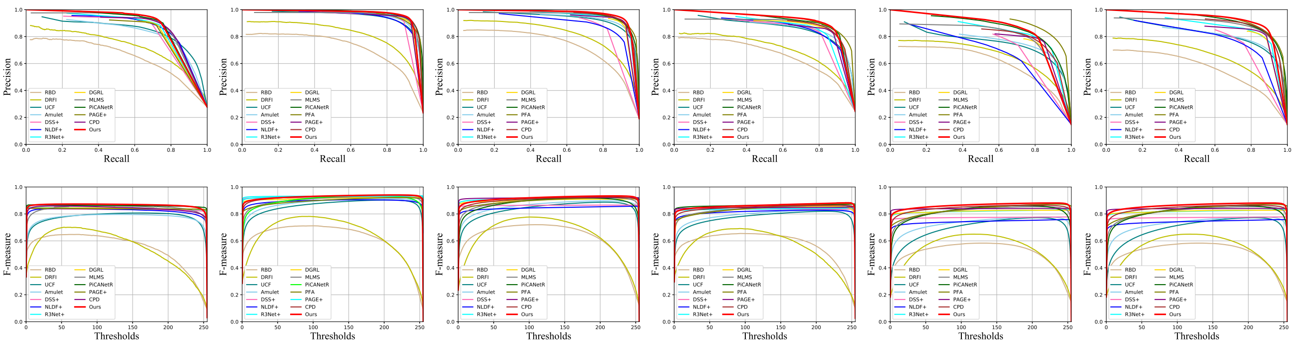
$$SAM = Softmax(conv_2(\sigma(conv_1(AvgPool(Z), W_1)), W_2))$$

$$Z' = SAM \odot Z$$

where $Z \in \mathbb{R}^{C \times H \times W}$ is a feature map, SAM is the output of the spatial attention module, AvgPool is the average pooling along

Table 3: F_β^{max} and MAE values for different saliency detection approaches on all the tested datasets. The two best results are marked in red and blue. "+" means that the results are generated with post-processing by CRF. "-" means that the author does not provide the saliency results on the dataset.

Methods	SOD		ECSSD		HKU-IS		PASCALS		DUT-OMRON		DUT-test	
	F_β^{max}	MAE	F_β^{max}	MAE	F_β^{max}	MAE	F_β^{max}	MAE	F_β^{max}	MAE	F_β^{max}	MAE
RBD [42]	0.648	0.228	0.712	0.172	0.720	0.142	0.654	0.193	0.628	0.142	0.583	0.152
DRFI [15]	0.701	0.223	0.782	0.170	0.777	0.144	0.691	0.196	0.664	0.150	0.649	0.154
UCF [39]	0.807	0.148	0.903	0.069	0.888	0.062	0.819	0.111	0.729	0.120	0.772	0.111
Amulet [38]	0.796	0.144	0.915	0.059	0.897	0.051	0.834	0.099	0.743	0.098	0.777	0.084
DSS+ [14]	0.845	0.122	0.921	0.052	0.866	0.059	0.836	0.102	0.745	0.075	0.778	0.069
NLDF+ [21]	0.840	0.123	0.905	0.063	0.858	0.060	0.828	0.101	0.679	0.107	0.758	0.077
R^3Net+ [9]	0.848	0.124	0.934	0.040	0.921	0.034	0.844	0.100	0.804	0.062	0.835	0.057
DGRL [28]	0.845	0.103	0.922	0.041	0.910	0.036	0.857	0.081	0.774	0.062	0.828	0.049
PiCANetR [20]	0.867	0.094	0.935	0.047	0.919	0.043	0.874	0.073	0.819	0.065	0.862	0.049
MLMS [31]	0.862	0.106	0.930	0.044	0.922	0.039	0.864	0.079	0.791	0.068	0.852	0.046
PFA [41]	-	-	0.922	0.045	0.931	0.032	0.871	0.077	0.862	0.058	0.872	0.039
PAGE+ [29]	0.842	0.108	0.934	0.037	0.921	0.031	0.853	0.083	0.794	0.059	0.841	0.047
CPD [32]	0.859	0.110	0.939	0.037	0.925	0.078	0.856	0.078	0.797	0.056	0.865	0.042
Ours	0.873	0.103	0.940	0.038	0.934	0.031	0.883	0.071	0.828	0.056	0.881	0.038



SOD ECSSD HKU-IS PASCALS DUT-OMRON DUT-test

Figure 5: Illustration of the PR curves (first row) and F-measure curves (last row) on the six widely used datasets.

the channel direction, $conv_1$ and $conv_2$ denote the convolution layer and batch normalization layer, which capture the spatial dependencies, W_1 and W_2 are the respective weights of $conv_1$ and $conv_2$, σ is the $Relu$ non-linear activation function, and $Softmax$ is used to enhance the most salient space and weaken the non-salient space. Finally, the weighted feature Z' is computed by performing the element multiplication between SAM and the original feature Z.

3.3 Loss Function

In machine learning and mathematical optimization, loss functions represent the cost of inaccurate predictions in classification problems. Same as [14], we use the cross-entropy loss between the final saliency map and the ground truth in saliency detection. The loss function is defined as

$$L_s = -\sum_{i=0}^{size(Y)} (Y_i \log(P_i) + (1 - Y_i \log(1 - P_i)))$$

where Y_i is the ground truth of pixel i, and P_i is the value of the predicted saliency map of pixel i.

4 Experimental Results

4.1 Datasets and Evaluation Metrics

4.1.1 Datasets

The performance evaluation was performed on six standard benchmark datasets: ECSSD [35], HKU-IS [17], SOD [22], PASCAL-S [18], DUT-OMRON [36] and DUTS-test [26]. ECSSD [35] contains 1000 images with many semantically meaningful and complex structures. HKU-IS [17] contains 4447 challenging images, each of which usually has multiple disconnected salient objects, overlapping the image boundary or low color contrast. SOD [22] includes 300 challenging images, which usually have complex backgrounds. PASCAL-S [18] contains 850 images selected from the PASCAL VOC 2010 segmentation dataset. DUT-OMRON [36] has 5168 high-quality images, which have one or more salient objects and relatively complex backgrounds. DUTS [26] is a large-scale dataset, which contains 10553 images for training and 5019 images for testing.

4.1.2 Evaluation Metrics

To quantitatively evaluate the improvements of the proposed model, we employed maximum F-measure, MAE scores and PR curve as the

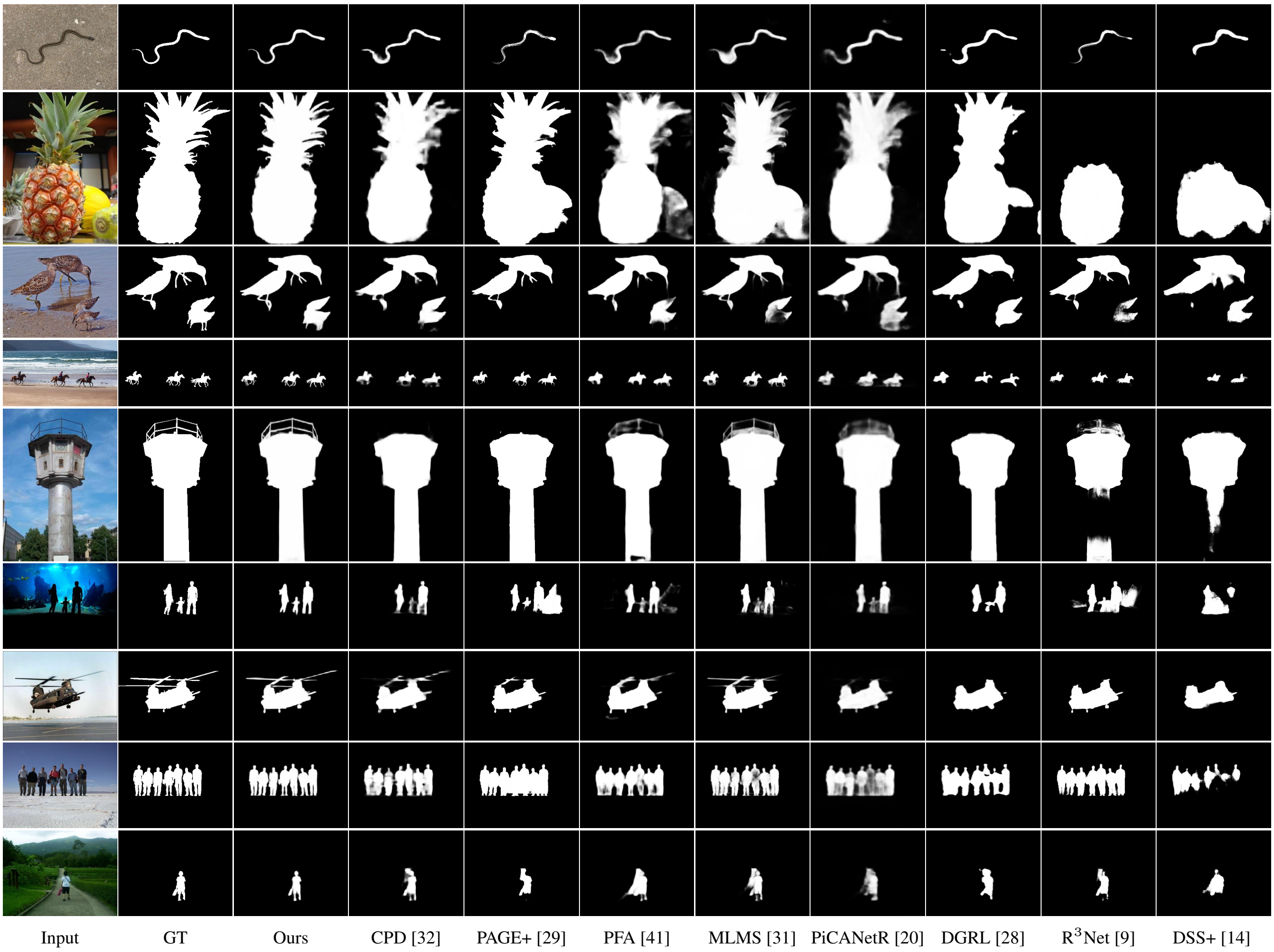


Figure 6: Visual comparison.

evaluation metrics. As described in [32], the metrics are computed as follows.

The precision of a binary map is defined as the ratio of the number of correctly labeled salient pixels to all the salient pixels in this binary map. The recall value is the ratio of the number of correctly labeled salient pixels to all the salient pixels in the ground-truth map. The formula is as follows,

$$precision = \frac{|TS \cap DS|}{|DS|}$$

$$recall = \frac{|TS \cap DS|}{|TS|}$$

where TS denotes the true salient pixels, DS denotes the salient pixels detected by the binary map, and $|\cdot|$ denotes the cardinality of a set.

Given a saliency map with continuous values normalized in the range of 0 to 255, we computed the corresponding binary maps by using every possible fixed integer threshold. Therefore, the F-measure curve can be obtained by connecting the F-measure scores under different thresholds. The maximum F-measure, denoted as F_β^{max}, is an overall performance indicator computed using the weighted harmonic of precision and recall,

$$F_\beta = \frac{(1+\beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$$

where β^2 is set as 0.3 to emphasize the precision, as suggested in [1].

The MAE is used to quantitatively measure the average difference between the saliency map of the network output P and the ground truth map Y.

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |P(x,y) - Y(x,y))|$$

The MAE value indicates the similarity of a saliency map compared to the ground truth [23].

The Precision-Recall (PR) curve is a standard metric to evaluate the saliency performance. The precision and recall are computed by comparing the predicted saliency map and the ground truth. Furthermore, the precision-recall pairs are computed considering all the saliency maps in a dataset under different thresholds, ranging from 0 to 255. These values are plotted as the PR curve.

4.2 Implementation Details

We used the Resnet-34 network [12] pre-trained on ImageNet [8] as our basic model. The DUTS-train dataset was used to train our model. As suggested in [19], we did not use the validation set and trained the model until the training loss converged. To make the model robust, we adopt several data augmentation techniques such as random brightness, saturation and contrast changing, and random horizontal flipping. In the training period, similar to other deep saliency methods [9], we used the stochastic gradient descent (SGD) to train the model, and setted the momentum as 0.9, weight decay as 0.0005, and learning rate as 0.001. We resized the input image to 256 x 256 for training, and the saliency map during testing was restored to the original size using bilinear interpolation. Our model was trained on a single 1080Ti GPU with a mini-batch size of 12, and it took about 12 hours to train the entire model. The inference for a 400×300 image took only 0.05s (20 fps) using the trained model.

4.3 Comparison with State-of-the-arts

We compared our method with 13 state-of-the-art approaches on six tested datasets, including CPD [32], PAGE [29], PFA [41], MLMS [31], PiCANet [20], DGRL [28], R^3Net [9], NLDF [21], DSS [14], Amulet [38], UCF [39], DRFI [15] and RBD [42]. For fair comparison, we use saliency maps provided by the authors or their released codes with default settings.

In Table 3, we show our quantitative comparison results. Some methods such as the DSS, R^3Net, and PAGE adopt the fully connected conditional random field (CRF [16]) as the post-processing to enhance the saliency map. Clearly, our model achieves the best results without any pre-processing and post-processing. In addition to the numerical comparisons, we plot the precision-recall curves and F-measure curves for all the compared methods over the six datasets. As shown in Figure 5, the solid red line, which represents the proposed method, corresponds to the best performance among all compared methods at most thresholds. In particular, the proposed approach exhibits the best performance among those of all the datasets in terms of the F-measure. Although the PFA method is superior to our method in terms of both the PR curve and the F_β^{max} on the DUT-OMRON dataset, our method is considerably more robust on datasets such as the ECSSD, PASCAL, and HKU-IS. These datasets are different from the DUT training set, and our method considerably outperforms the PFA on these datasets.

In Figure 6, we show the qualitative comparison. It can be observed that the proposed model can handle various challenging scenarios, including images with low contrast (rows 1, 5, and 9), complex object boundaries (rows 2 and 5), varying object scales (rows 3 and 6), small scale objects (rows 4 and 9), objects touching the image boundary (row 5) and multiple objects (rows 4, 6 and 8).

4.4 Ablation Study

To investigate the importance of the different modules in our method, we conducted an ablation study as shown in Table 4, where a higher F_β^{max}, and lower MAE correspond to better results. The proposed model containing all the components (i.e., the Basic Resnet-34 (BASIC), Deformable Feature Extraction Module (DFEM), Channel Attention Module (CAM) and Spatial Attention Module (SAM)) achieves the best performance. This demonstrates that all the components are necessary for the proposed method to obtain the best salient object detection result.

5 Conclusion

This paper proposes a novel saliency detection model named DCFA. We design a deformable feature extraction module to capture unevenly distributed features to improve the saliency detection results. Furthermore, we employ a channel and spatial attention module to focus on the crucial features and suppress the noises. The proposed model achieves excellent performance and produces visually favorable results. The experimental results on six widely used datasets verify that our proposed approach outperforms 13 other state-of-the-art methods under different evaluation metrics. Besides, the proposed method is an end-to-end network and runs at a speed of 20 FPS in the inference period.

Acknowledgements. This work is partially supported by the National Natural Science Foundation of China (No.61661146002).

Table 4: Ablation study using different component combinations.

BASIC	DFEM	CAM	SAM	F_β^{max}	MAE
✓				0.928	0.049
✓	✓			0.934	0.042
✓	✓	✓		0.937	0.040
✓	✓		✓	0.938	0.040
✓	✓	✓	✓	0.940	0.038

REFERENCES

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk, 'Frequency-tuned salient region detection', in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, number CONF, pp. 1597–1604, (2009).

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, 'Neural machine translation by jointly learning to align and translate', *arXiv preprint arXiv:1409.0473*, (2014).

[3] Sai Bi, Guanbin Li, and Yizhou Yu, 'Person re-identification using multiple experts with random subspaces', *Journal of Image and Graphics*, **2**(2), (2014).

[4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, 'Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5659–5667, (2017).

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

[6] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang, 'Multi-context attention for human pose estimation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1831–1840, (2017).

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei, 'Deformable convolutional networks', in *Proceedings of the IEEE international conference on computer vision*, pp. 764–773, (2017).

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, 'Imagenet: A large-scale hierarchical image database', in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, (2009).

[9] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng, 'R3net: Recurrent residual refinement network for saliency detection', in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 684–690. AAAI Press, (2018).

[10] Michael Donoser, Martin Urschler, Martin Hirzer, and Horst Bischof, 'Saliency driven total variation segmentation', in *2009 IEEE 12th International Conference on Computer Vision*, pp. 817–824, (2009).

[11] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin, 'Convolutional sequence to sequence learning', in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1243–1252. JMLR. org, (2017).

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, (2016).

[13] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han, 'Online tracking by learning discriminative saliency map with convolu-

tional neural network', in *International conference on machine learning*, pp. 597–606, (2015).

[14] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr, 'Deeply supervised salient object detection with short connections', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3203–3212, (2017).

[15] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li, 'Salient object detection: A discriminative regional feature integration approach', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2083–2090, (2013).

[16] Philipp Krähenbühl and Vladlen Koltun, 'Efficient inference in fully connected crfs with gaussian edge potentials', in *Advances in neural information processing systems*, pp. 109–117, (2011).

[17] Guanbin Li and Yizhou Yu, 'Visual saliency based on multiscale deep features', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5455–5463, (2015).

[18] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille, 'The secrets of salient object segmentation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280–287, (2014).

[19] Nian Liu and Junwei Han, 'Dhsnet: Deep hierarchical saliency network for salient object detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 678–686, (2016).

[20] Nian Liu, Junwei Han, and Ming-Hsuan Yang, 'Picanet: Learning pixel-wise contextual attention for saliency detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3089–3098, (2018).

[21] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin, 'Non-local deep features for salient object detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6609–6617, (2017).

[22] Vida Movahedi and James H Elder, 'Design and perceptual validation of performance measures for salient object segmentation', in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pp. 49–56. IEEE, (2010).

[23] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung, 'Saliency filters: Contrast based filtering for salient region detection', in *2012 IEEE conference on computer vision and pattern recognition*, pp. 733–740. IEEE, (2012).

[24] Pierre Sermanet, Andrea Frome, and Esteban Real, 'Attention for fine-grained categorization', *arXiv preprint arXiv:1412.7054*, (2014).

[25] 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 Conference on Computer Vision and Pattern Recognition*, pp. 3183–3192, (2015).

[26] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan, 'Learning to detect salient objects with image-level supervision', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 136–145, (2017).

[27] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan, 'Saliency detection with recurrent fully convolutional networks', in *European conference on computer vision*, pp. 825–841. Springer, (2016).

[28] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji, 'Detect globally, refine locally: A novel approach to saliency detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3127–3135, (2018).

[29] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji, 'Salient object detection with pyramid attention and salient edges', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1448–1457, (2019).

[30] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun, 'Geodesic saliency using background priors', in *European conference on computer vision*, pp. 29–42. Springer, (2012).

[31] Runmin Wu, Mengyang Feng, Wenlong Guan, Dong Wang, Huchuan Lu, and Errui Ding, 'A mutual learning method for salient object detection with intertwined multi-supervision', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8150–8159, (2019).

[32] Zhe Wu, Li Su, and Qingming Huang, 'Cascaded partial decoder for fast and accurate salient object detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3907–3916, (2019).

[33] Saining Xie and Zhuowen Tu, 'Holistically-nested edge detection', in *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, (2015).

[34] Huijuan Xu and Kate Saenko, 'Ask, attend and answer: Exploring question-guided spatial attention for visual question answering', in *European Conference on Computer Vision*, pp. 451–466. Springer, (2016).

[35] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia, 'Hierarchical saliency detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1155–1162, (2013).

[36] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang, 'Saliency detection via graph-based manifold ranking', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3166–3173, (2013).

[37] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola, 'Stacked attention networks for image question answering', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, (2016).

[38] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan, 'Amulet: Aggregating multi-level convolutional features for salient object detection', in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 202–211, (2017).

[39] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin, 'Learning uncertain convolutional features for accurate saliency detection', in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 212–221, (2017).

[40] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang, 'Progressive attention guided recurrent network for salient object detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 714–722, (2018).

[41] Ting Zhao and Xiangqian Wu, 'Pyramid feature attention network for saliency detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3085–3094, (2019).

[42] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun, 'Saliency optimization from robust background detection', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2814–2821, (2014).