

Extended non-local feature for visual saliency detection in low contrast images

Xin Xu^{1,2,*}, Jie Wang¹

¹ School of Computer Science and Technology, Wuhan University of Science and Technology, 430065 Wuhan, China

² Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, 430065 Wuhan, China
xuxin0336@163.com

Abstract. Saliency detection model can substantially facilitate a wide range of applications. Conventional saliency detection models primarily rely on high level features from deep learning and hand-crafted low-level image features. However, they may face great challenges in nighttime scenario, due to the lack of well-defined feature to represent saliency information in low contrast images. This paper proposes a saliency detection model for nighttime scene. This model is capable of extracting non-local feature that is jointly learned with local features under a unified deep learning framework. The key idea of the proposed model is to hierarchically introduce non-local module with local contrast processing blocks, aiming to provide robust representation of saliency information towards low contrast images with low signal-to-noise ratio property. Besides, both nighttime and daytime images are utilized in training to provide complementary information. Extensive experiments have been conducted on five challenging datasets and our nighttime image dataset to evaluate the performance of the proposed model.

Keywords: deep learning, non-local feature, saliency detection, low contrast images.

1 Introduction

The purpose of saliency detection is to highlight significant areas and targets in images. Saliency detection aims to mimic the human visual system, which can naturally separate predominant objects of a scene from the rest of image. As a computer vision preprocessing step, saliency detection has achieved great success in various applications, such as object retargeting [1], photo synthesis [2-3], visual tracking [4], image retrieval [5-6], semantic segmentation [7], and etc.

Conventional saliency detection models primarily extract effective information from images based on low-level visual features [8-10]. With the development of deep learning in recent years, high level features extracted from deep learning have demonstrated superior results in saliency detection. Current deep learning based saliency detection models can be generally divided into three categories: 1. Global features extraction using *convolutional neural network* (CNN); 2. Multi-scale local

features extraction; and 3. Constructing non-local neural networks to integrate global and local features. The first type extracts global features containing image objectiveness via a straight forward CNN model [11-13]. The second one extracts local image features incorporating multi-task processing, such as generative object proposals, post-processing, superpixel smoothing, superpixel segmentation [1, 14-17], and etc. However, either global features or local features can only reflect partial aspect of visual saliency and may cause certain bias. Combination of the information from both global and local features can be accurate and effective. Accordingly, the third category utilizes non-local structure to extract local and global features [18-19]. The non-local structure has demonstrated its effectiveness efficiency in saliency detection.

However, current non-local based saliency detection models simply perform mean processing or short connection to different feature layers. They are mainly based on patch operation, and may face great challenges in nighttime scenario, due to the lack of well-defined feature to represent saliency information in low contrast images. In this paper, we propose a novel saliency detection model for nighttime scene. This model can extract non-local feature that is jointly learned with local features under a unified deep learning framework.

The rest of the paper is organized as follows. Section 2 provides an overview of saliency detection models. Section 3 describes the theory and practical implementation of our network. Section 4 shows the performance of the proposed model against the state-of-the-art models. Finally, Section 5 gives the conclusions.

2 Related works

Most of current saliency detection models highlight salient object by comparing its difference with backgrounds, and primarily rely on low level features, including color [8], contrast [9], contour [10], objectness [20], focusness [21], backgroundness [22], uniqueness [23], and etc. These methods do not need the training process, and extract saliency feature at pixel level [9], region level [8] and graph [2] respectively. Recently, deep learning models have demonstrated their effectiveness in saliency detection, which can extract high level features directly from image.

Early deep learning based saliency detection models [12-13] mainly utilize convolutional layer to obtain the global features in images, and use fully connected layers for output. However, this structure only extracts objectiveness features in images, and can only roughly determine the location of salient object with incomplete information. Aiming to address this problem, local neural networks [1] and multi-tasking neural networks [16] are proposed recently. For example, Li *et al.* [1] proposed the *multiscale deep features* (MSDF) neural network, which decomposes input images into a set of non-overlapping blocks and then puts them into the three-scale neural networks to learn the local features, finally outputs with a full connected layer. Zhao *et al.* [16] proposed the *multi-task* (MT) neural network, which uses convolution to extract global feeble features and combines superpixel segmentation to jointly guide the output of saliency maps.

However, multiple levels of convolutional and pooling layers “blur” the object boundaries, and high level features from the output of the last layer are too coarse spatially for the saliency detection task. Accordingly, the non-local neural networks [18-19] are proposed to improve the performance. Luo *et al.* [18] proposed the *non-local deep features* (NLDF) network, which uses the convolution to extract local and global features. Then it uses upsampling to connect each local feature. Finally, the local and global features are linearly fused to output the saliency map. In order to get the local depth feature, it subtracts the local mean from the local feature in the contrast layer, so that a simple processing is done on the pixel-wise. Chen *et al.* [19] proposed a *deeply supervised short connections* (DSSC) neural network by upsampling to connect low-level and high-level features shortly, so that high-level features can share the information from the low-level features. Both of these methods increase the receptive fields of convolution, and greatly improve their effectiveness to avoid blurring object boundaries.

However, current non-local based saliency detection models simply perform mean processing or short connection to different feature layers [24]. They are mainly based on patch operation, and may face great challenges in nighttime scenario, due to the lack of well-defined feature to represent saliency information in low contrast images. In this paper, we propose a novel saliency detection model for nighttime scene. As illustrated in Fig. 1, our model differs from current models as it extracts non-local feature which is jointly learned with local features under a unified deep learning framework.

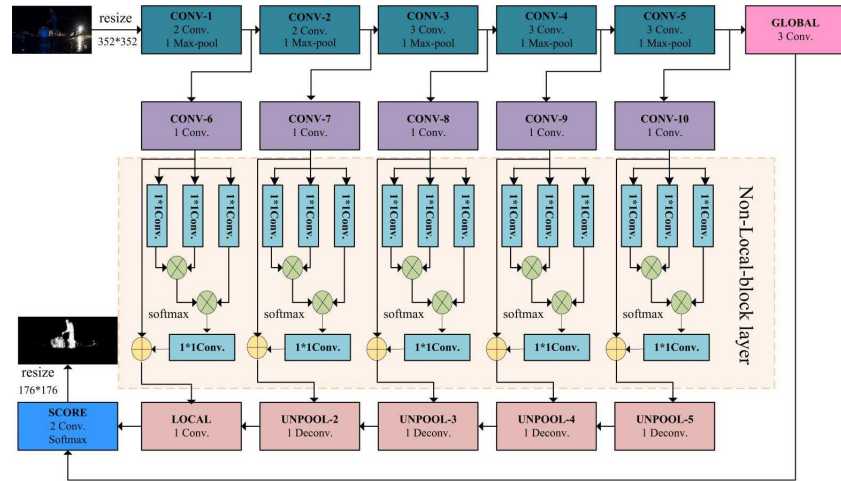


Fig. 1. Architecture of our 4×5 grid-CNN network for salient object detection.

The main attributions of the proposed model are in three folds:

1. The model employs non-local blocks with local contrast processing units to learn saliency information from low contrast images;
2. The model introduces an IoU boundary loss to the loss function to make the boundary robust in training process;

3. Both nighttime and daytime images are used in training. Although the proposed model still falls behind the existing deep saliency models on daytime images, it receives the highest performance on nighttime images. Thus, the experimental results show that the non-local block layers efficiently extract local details on low contrast images.

3 Proposed Model

3.1 Network Architecture

As illustrated in Fig. 1, this paper provides a deep convolutional network architecture to learn discriminant saliency features from nighttime scene. Both local and global features are incorporated for salient object detection. In additions, pixel-wise calculating can provide sufficient information from low contrast images. Specifically, we have implemented a novel grid-like CNN network containing 5 columns and 4 rows. Each column extracts features at a given input scale. As illustrated in Fig. 2, the input image (on the left) is a 352×352 image and the output (on the right) is a 176×176 saliency map which was resized to 352×352 via bilinear interpolation.

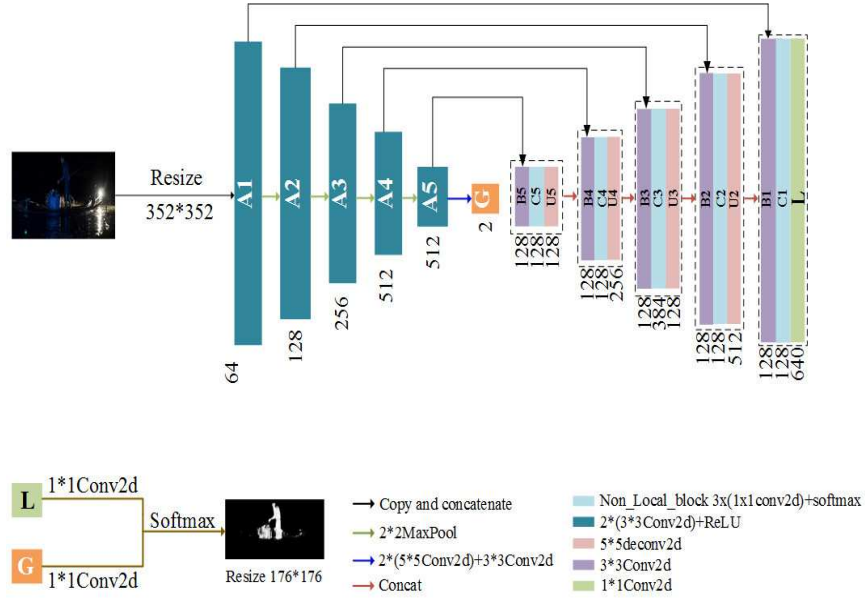


Fig. 2. Network: As an input, we have RGB channels image. A1-A5 feature maps are obtained by the first layer with five convolutional blocks. The global (G) feature map is acquired after A5. B1-B5 are computed by the second layer with five convolutional blocks that change the channels to 128. C1-C5 are calculated by the third layer with five non-local blocks that obtain more useful features from low contrast images .we perform upsampling on last layer which generated U2-U5, followed by the series of the deconvolution layers. A 1x1 convolution is added after C1 to sum the number of channels to 640, and then local feature map L is gained. Finally, G and L are liner-fused by a 1x1 convolution to generate the saliency map.

The first row of our model contains five convolutional blocks derived from VGG-16 [1] (CONV-1 to CONV-5), as shown in Fig. 1. These convolution blocks contain a max pooling operation of stride 2 which downsamples their feature maps $\{A1, K, A5\}$, as shown in Fig. 2. The last and rightmost convolution block of the first row computes features G that are specific to the global context of the image.

The second and third row is a set of ten convolutional blocks, CONV-6 to CONV-10 for row 2 and non-local layer for row 3(see in Fig. 1). The aim of these blocks is to compute the similarity of any two pixels by self-attention to each resolution. The non-local layer capture the difference of each feature against its local neighborhood favoring regions that are either brighter or darker than their neighbors

The last row is a set of deconvolution layers used to upscale the features maps from 11×11 (bottom right) all the way to 176×176 (bottom left). These UNPOOL layers are a means of combining the feature maps (C_i, U_i) computed at each scale. The lower left block constructs the final local feature map L . The SCORE block has 2 convolution layers and a softmax to compute the saliency probability by fusing the local (L) and global (G) features. Further details of our model are given in Fig. 2.

3.2 Non-Local Feature Extraction

First, the size of input images is resized to 352×352 , and then the feature maps of the first layer in the network are extracted by VGG-16 (Conv-1 to Conv-5), denoted as A_i , $i = 1, L, 5$. Finally, the feature maps outputted by VGG-16 are connected by the convolutional blocks CONV-6 to CONV-10 each of which has a kernel with size 3×3 and 128 channels. The feature maps after the convolution are denoted as B_i , $i = 1, L, 5$.

In the architecture of NLDF, the contrast features layer adopts a simple mean layer, which cannot obtain a larger receptive field in local features. Differently, in this layer, we use the architecture of non-local block to generate three feature maps by 1×1 convolution of the input value B_i . Next, the similarity of any two pixels in the feature map is determined by Gaussian filter, which makes up for the lack of local computing information of a single mean layer. At last, the weight of each pixel in the feature map is updated by residual network, so that the salient object in the feature map is more prominent to achieve the purpose of noise reduction, and acquire more useful features from low contrast images, and make the edge of the salient object clearer.

In order to learn more useful information from low contrast images, we are motivated by non-local mean [25] and bilateral filters [26], and then take advantage of

the matrix multiplication to calculate the similarity of any two pixels and make the feature map embedded into Gaussian after 1×1 convolution, which is defined as:

$$f(x_i, x_j) = e^{(W_\theta x_i)^T W_\phi x_j}, \quad (1)$$

where x_i, x_j represent any two pixels of Bi . W_θ and W_ϕ represent the weight of convolution. After the convolution, the number of channels becomes half as many as it was before.

The similarity calculated above is stored in feature maps by means of self-attention, which is defined by $y_i = \text{soft max}(Bi^T W_\theta^T W_\phi Bi)g(Bi)$. After that, the feature map Ci , $i = 1, L, 5$ is obtained through a process of residual operation by y_i and Bi via:

$$Ci = W_B y_i + Bi, \quad (2)$$

where W_B is a weighting parameter to restore the same number of channels y_i same as Bi . Therefore, the size of the Ci feature map is the same as before after the process of the non-local network layer Bi .

The last layer is the deconvolution layer, which is designed to connect the precomputed local features of the five branches of network inversely one by one. At the same time, each size of the feature map is increased by a ratio of $\{2, 4, 8, 16\}$. By doing so, the information expressed by the feature map becomes richer. Different from the NLDF [18], we replaced the mean layer with a non-local module layer, the output of which is connected by upsampling. The feature map deconvolved is defined as $Ui = \text{UNPOOL}(Ci, U(i+1))$, where the Ui , $i = 2, K, 5$ is the resulting unpooled feature map. After that, the local feature map (denoted as L) is acquired by.

$$L = \text{CONV}(C1, U2) \quad (3)$$

3.3 Cross Entropy Loss

We adopt the method of linear combination to combine the local features L and global features G .

$$\hat{y}(v) = p(y(v) = c) = \frac{e^{W_L^c L(v) + b_L^c + W_G^c G + b_G^c}}{\sum_{c' \in \{0,1\}} e^{W_L^{c'} L(v) + b_L^{c'} + W_G^{c'} G + b_G^{c'}}}, \quad (4)$$

The formula uses two linear operators (W_L, b_L) and (W_G, b_G) . The $y(v)$ represents ground truth. The final saliency map is predicted as $\hat{y}(v_i)$.

The cross-entropy loss function is defined as:

$$H_j(y(v), \hat{y}(v)) = -\frac{1}{N} \sum_{i=1}^N \sum_{c \in \{0,1\}} (y(v_i) = c)(\log(\hat{y}(v_i) = c)). \quad (5)$$

What's more, we make great use of the IoU boundary loss of NLDF [18] to make the boundary robust.

$$IoU(C_j, \hat{C}_j) = 1 - \frac{2|C_j \cap \hat{C}_j|}{|C_j| + |\hat{C}_j|}. \quad (6)$$

Finally, the final loss function is obtained by a combination of the cross-entropy loss function and the IoU boundary loss,

$$Total\ Loss \approx \sum_j \lambda_j \int H_j(y(v), \hat{y}(v)) + \sum_j \gamma_j (1 - IoU(C_j, \hat{C}_j)). \quad (7)$$

Our whole loss computation procedure is end-to-end train, and an example is shown in Fig. 3.

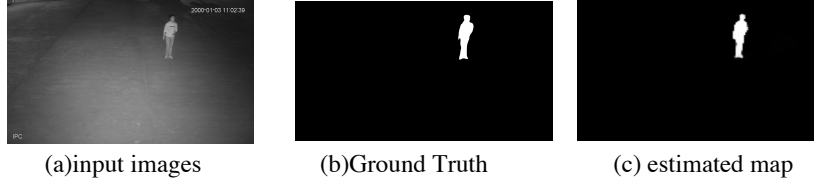


Fig. 3. A single input image. (a) together with its ground truth saliency; (b) the estimated boundary; (c) after training for 17 epochs is in good agreement with the true bound.

4 Experiments

4.1 Datasets

In order to evaluate the performance of the proposed approach, we conduct a set of qualitative and quantitative experiments on six benchmark datasets annotated with pixelwise ground-truth labeling, including MSRA-B [27], HKU-IS [1], DUT-OMRON [28], PASCAL-S [29], and ECSSD [30]. Besides, we built a nighttime images (NTI) dataset with 478 nighttime natural scene images to facilitate this study.

MSRA-B: contains 5000 images, most of which have one salient object and corresponding pixel ground truth [31].

HKU-IS: contains 4447 images, most of which are used for multiple salient objects.

DUT-OMRON: contains 5168 images, each of which contains one or more new salient objects with a complex background.

PASCAL-S: contains 850 images. This dataset contains both pixel-wise saliency ground truth and eye fixation ground truth labeled by 12 subjects.

ECSSD: contains 1000 images with complicated architecture all of which are collected from the Internet. The ground truth masks were labeled by 5 subjects.

NTI: contains 478 nighttime natural scene images. This dataset contains two degree low contrast images, which consists of 3 subjects, the one about Only a person, the another with many people, others included human and object(such as bicycle, car, and house and etc.). So the model with low contrast features can be learned via the dataset.

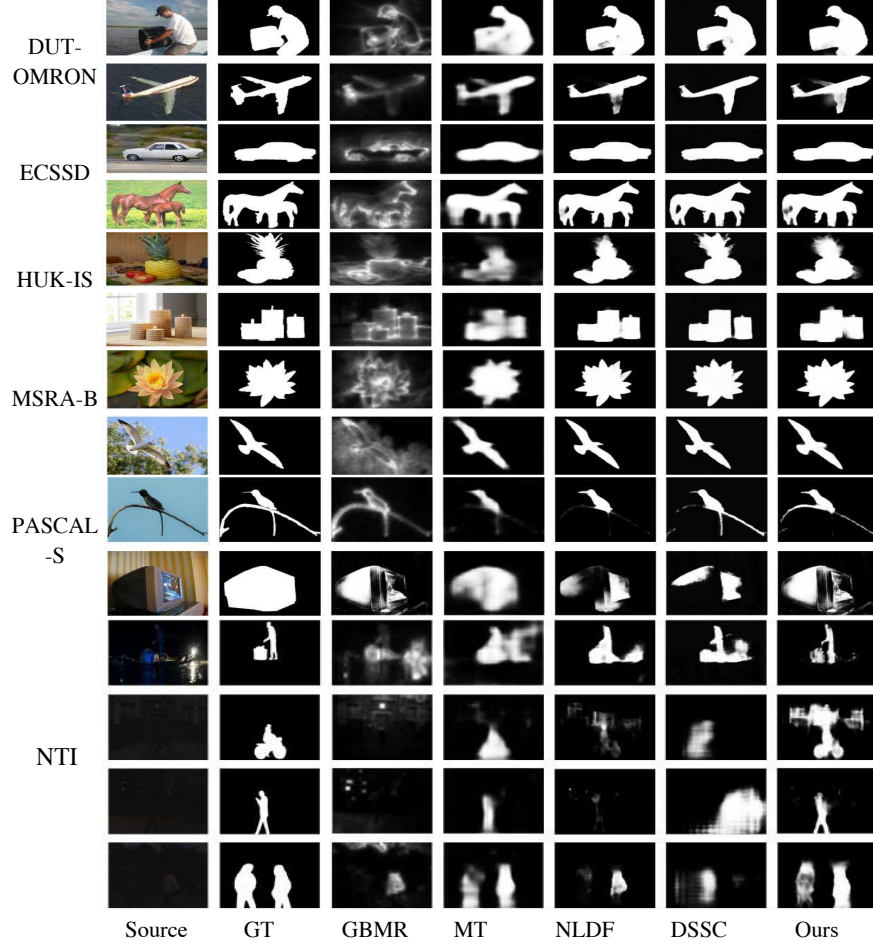


Fig. 4. Saliency maps produced by the GBMR[32], MT [16], DSSC[19], NLDF[18] methods compared to our method on six datasets. The Our maps provides clear salient regions and exhibit good uniformity as compared to the saliency maps from the other deep learning methods (MT, NLDF, DSSC) on NTI dataset. Our method is also more robust to background clutter than the non-deep-learning method (GBMR).

4.2 Implementation and Experimental Setup

Our method is accomplished by TensorFlow [1]. The weights of CONV-1 to CONV-5 are initialized with network of VGG-16 [13]. All of the weights added in the network were initialized randomly by a truncated normal ($\sigma = 0.01$). Besides, the biases were initialized to zero. There is an adam optimizer [33] used to train our model with a learning rate of 10^{-6} , $\beta_1 = 0.9$, and $\beta_2 = 0.999$.

In our experiment, the datasets of MSRA-B and NTI were divided into three parts: the 1000 images in MSRA-B and 220 images in NTI were used to train, and the validation set included 500 images in MSRA-B and 100 images in NTI, the rest of which were added to the test set. Our models were trained by the combination of training set and validation set. What's more, the method of horizontal flipping is adopted to achieve the purpose of data augmentation. The inputs were resized to 352×352 for the training of network. It takes about seven hours for 17 epochs in the configuration of NVIDIA 1070.

4.3 Evaluation Criteria

In this paper, we make use of *precision-recall* (PR) curves, F_β and *mean absolute error* (MAE) to evaluate the performance of saliency detection. By binarizing the saliency maps with different thresholds which range from 0 to 1 and comparing against the ground truth, the PR curve is obtained. The F_β is defined as,

$$F_\beta = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}, \quad (8)$$

where β^2 is valued by 0.3 as usual so that the precision over recall can be emphasized just like [34]. The maximum F-Measure is computed from the PR curve. The MAE [35] is defined as

$$MAE = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H |S(x, y) - L(x, y)|, \quad (9)$$

where the function of $S(x, y)$ is a predicted salient map and $L(x, y)$ is the ground truth. The parameters of W and H represent the width and height, respectively.

4.4 Data fusion

There are three models obtained by the progress of training. We call the model trained with only the night images NT-model and the model trained with high contrast images is called the DT-model, Furthermore, the NDT-model was defined by a model trained by combining night images with high contrast images. The performance of the models is shown in Fig. 5. MAE and Max F_β are illustrated in Table. 1.

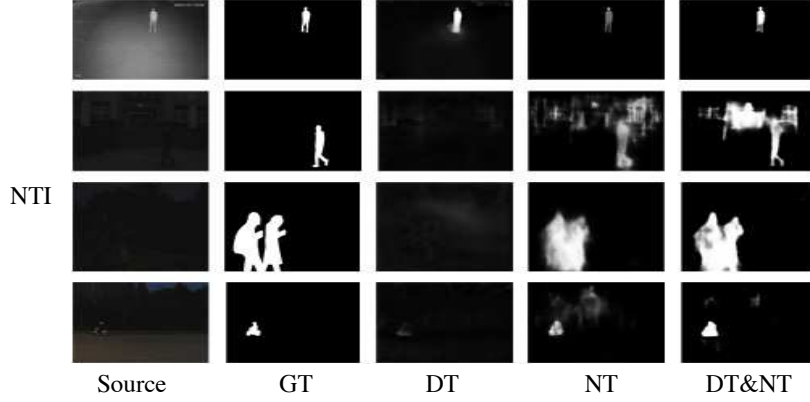


Fig. 5. Daytime datasets to get the model DT and the model NT obtained by the night datasets. Naturally we acquired the model of DT&NT with a fusion of them. It turned out that the DT detected no objects and NT measured objects with relatively blurry edges. After a fusion of them, the performance of the model is greatly improved.

Through the test, we can see from the evaluation indicators that the NDT-model is 13.1% lower than the DT-model in MAE, and a 46.5% increase of $\text{Max } F_\beta$. In addition, the NDT-model is 11% higher than the NT-model, and the MAE is 1% lower (see Table. 1). The model after a data fusion becomes more robust than before (see Fig. 5).

Table 1. MAE and $\text{Max } F_\beta$ performance of NT-model, DT-model, NDT-model.

Dataset	Metric	Daytime (DT)	Nighttime (NT)	DT&NT(NDT)
NTI	$\text{Max } F_\beta$	0.316	0.631	0.741
	MAE	0.171	0.050	0.040

4.5 Comparison with the State-of-the-Art

Visual comparison of the saliency maps is provided in Fig. 4. All saliency maps of other methods were either provided by the authors or computed using the authors' released code. PR curves are shown in Fig. 6, and the $\text{Max } F_\beta$ and MAE scores are in Table 2.

Table 2. Quantitative performance of our model on six benchmark datasets compared with the GBMR [32], MT [16], DSSC [19], and NLDF [18] models. The latter three are deep learning methods and the former is not. The $\text{Max } F_\beta$ and MAE metrics are defined in the text.

Dataset	Metric	GBMR	MT	NLDF	SC	ours
DUT-	$\text{Max } F_\beta$	0.474	0.774	0.753	0.726	0.747

OMRON	MAE	0.247	0.084	0.080	0.113	0.088
	Max F_β	0.549	0.900	0.905	0.914	0.896
ECSSD	MAE	0.297	0.079	0.063	0.069	0.066
	Max F_β	0.525	0.871	0.915	0.928	0.907
HKU-IS	MAE	0.267	0.084	0.049	0.069	0.052
	Max F_β	0.592	0.893	0.941	0.884	0.903
MSRA-B	MAE	0.245	0.069	0.030	0.075	0.049
PASCAL-S	Max F_β	0.587	0.856	0.849	0.851	0.835
	MAE	0.320	0.140	0.146	0.148	0.149
	Max F_β	0.271	0.667	0.668	0.423	0.741
NTI	MAE	0.080	0.048	0.042	0.135	0.040

Our network structure is similar to NLDF. Differently, a non-local block is added into the local module to calculate any two pixels similarity of the feature maps by self-attention. In result, the MAE decreased by 0.2%, Max F_β increased by 7.3% compared with NLDF in NTI dataset.

Although low-level and high-level features are combined by short connections to make the feature map more informative in DSSC, it is difficult to learn some useful features for the night scene. Thus, more useful features are obtained via non-local block for low contrast images in our method.

Moreover, the MT adopted superpixel segmentation to enhance the correlation between pixels in the environment of low SNR, but the convolution model is too simple to learn serviceable features. We took great advantage of the non-local network to compute the similarity of any two pixels for a better effect in NTI dataset.

As for the traditional method GBMR, it is difficult to find an effective feature applying to nighttime scenes. Differently, the proposed model adopted a data-driven approach to gain more effective features to make our method more robust.

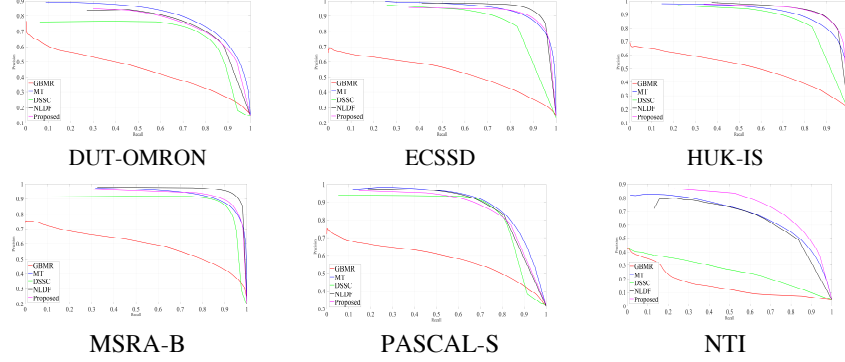


Fig. 6. PR curves for our model compared to GBMR [32], MT [16], DCSS [19], and NLDF [18]. Our model can deliver state-of-the-art performance on NTI datasets.

Since our method is designed for nighttime scenes, the daytime images can be optionally used for data fusion to improve the performance at nighttime. As illustrated in Fig. 6, the proposed model can achieve the best performance compared to NLDF, MT, DSSC and GBMR.

5 Conclusion

In this paper, we utilized a unified deep learning framework to integrate local and global features, and introduce non-local module with local contrast processing blocks. This method can provide robust representation of saliency information towards low contrast images with low signal-to-noise ratio property. Moreover, we utilize both nighttime and daytime images in training, which can provide complementary information to enhance performance of saliency detection. Our method has achieved the best performance compared to the state-of-the-art methods.

References

1. Li, G., and Yu, Y.: Visual saliency based on multiscale deep features. in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 5455--5463. (2015)
2. Chen, T., Cheng, M.-M., Tan, P., Shamir, A., and Hu, S.-M.: Sketch2photo: Internet image montage. *ACM Transactions on Graphics*, vol. 28, no. 5, pp.124:1--124:10. (2009)
3. Hu, S.-M., Chen, T., Xu, K., Cheng, M.-M., and Martin, R.-R.: Internet visual media processing: a survey with graphics and vision applications. *Visual Compute*, vol. 29, no. 5, pp. 393--405. (2013)
4. Borji, A., Frintrop, S., Sihite, D.-N., and Itti, L.: Adaptive object tracking by learning background context. in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 23--30. (2012)
5. Gao, Y., Wang, M., Tao, D., Ji, R., and Dai, Q.: 3-D object retrieval and recognition with hypergraph analysis. *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4290--4303. (2012)
6. Cheng, M.-M., Hou, Q.-B., Zhang, S.-H., and Rosin, P.-L.: Intelligent visual media processing: When graphics meets vision. *Journal of Computer Science and Technology*, vol. 32, no. 1, pp. 110--121. (2017)
7. Mehrani, P., and Veksler, O.: Saliency segmentation based on learning and graph cut refinement. in: British Machine Vision Conference, pp. 110.1--110.12. (2010)
8. Zhang, J., Wang, M., Zhang, S., Li, X., and Wu, X.: Spatiochromatic context modeling for color saliency analysis. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1177--1189. (2016)
9. Goferman, S., Zelnik-Manor, L., and Tal, A.: Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915--1926. (2012)
10. Liu, Q., Hong, X., Zou, B., Chen, J., Chen, Z., and Zhao, G.: Hierarchical contour closure-based holistic salient object detection. *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4537--4552. (2017)
11. Mu, N., Xu, X., Zhang, X., and Zhang, H.: Salient object detection using a covariance-based CNN model in low-contrast images. *Neural Computing and Applications*, vol. 29, no. 8, pp. 181--192. (2018)

12. Pan, J., Sayrol, E., Giro-I-Nieto, X., McGuinness, K., O'Connor, N.-E.: Shallow and deep convolutional networks for saliency prediction. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 598--606. (2016)
13. Simonyan, K., and Zisserman, A.: Very deep convolutional networks for large-scale image recognition. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1--14. (2014)
14. Li, G., and Yu, Y.: Deep contrast learning for salient object detection. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 478-487. (2016)
15. Liu, N., Han, J., Zhang, D., Wen, S., and Liu, T.: Predicting eye fixations using convolutional neural networks. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 362--370. (2015)
16. Li, X., Zhao, L., Wei, L., Yang, M.-H., Wu, F., Zhuang, Y., Ling, H., and Wang, J.: DeepSaliency: Multi-task deep neural network model for salient object detection. IEEE Transactions on Image Processing, vol. 25, no. 8, pp. 3919--3930. (2016)
17. Zhao, R., Ouyang, W., Li, H., and Wang, X.: Saliency detection by multi-context deep learning. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1265--1274. (2015)
18. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.-M.: Non-local deep features for salient object detection. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6593--6601. (2017)
19. Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5300--5309. (2017)
20. Jiang, P., Ling, H., Yu, J., and Peng, J.: Salient region detection by ufo: Uniqueness, focusness and objectness. in: IEEE International Conference on Computer Vision, pp. 1976--1983. (2013)
21. Wei, Y., Wen, F., Zhu, W., and Sun, J.: Geodesic saliency using background priors. in: European Conference on Computer Vision, pp. 29--42. (2012)
22. Chang, K.-Y., Liu, T.-L., Chen, H.-T., and Lai, S.-H.: Fusing generic objectness and visual saliency for salient object detection. in: IEEE International Conference on Computer Vision, pp. 914--921. (2011)
23. Perazzi, F., Krahenbuhl, P., Pritch, Y., and Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: IEEE International Conference on Computer Vision, pp. 733--740. (2012)
24. Xiao, L., Girshick, R., Gupta, A., Kaiming, H.: Non-local Neural Networks, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794--7803. (2018)
25. Buades, A., Coll, B., and Morel, J.-M.: A non-local algorithm for image denoising. in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 60--65. (2005)
26. Tomasi, C., and Manduchi, R.: Bilateral filtering for gray and color images. in: IEEE International Conference on Computer Vision, pp. 839--846. (1998)
27. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., and Shum, H.: Learning to detect a salient object. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 2, pp. 353--367. (2011)
28. Yang, C., Zhang, L., Lu, H., Ruan, X., and Yang, M.: Saliency detection via graph-based manifold ranking. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166--3173. (2013)
29. Li, Y., Hou, X., Koch, C., Rehg, J., and Yuille, A.: The secrets of salient object segmentation. in: IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 280--287. (2014)
30. Yan, Q., Xu, L., Shi, J., and Jia, J.: Hierarchical saliency detection. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1155--1162. (2013)

31. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., and Li, S.: Salient object detection: A discriminative regional feature integration approach. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2083--2090. (2013)
32. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.: Saliency detection via graph-based manifold ranking. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166--3173. (2013)
33. Kingma, D. P., and Ba, J.: Adam: A method for stochastic optimization. in: International Conference for Learning Representations, pp. 1--15. (2014)
34. Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S.: Frequency-tuned salient region detection. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597--1604. (2009)
35. Perazzi, F., Krahenbuhl, P., Pritch, Y., and Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 733--740. (2012)